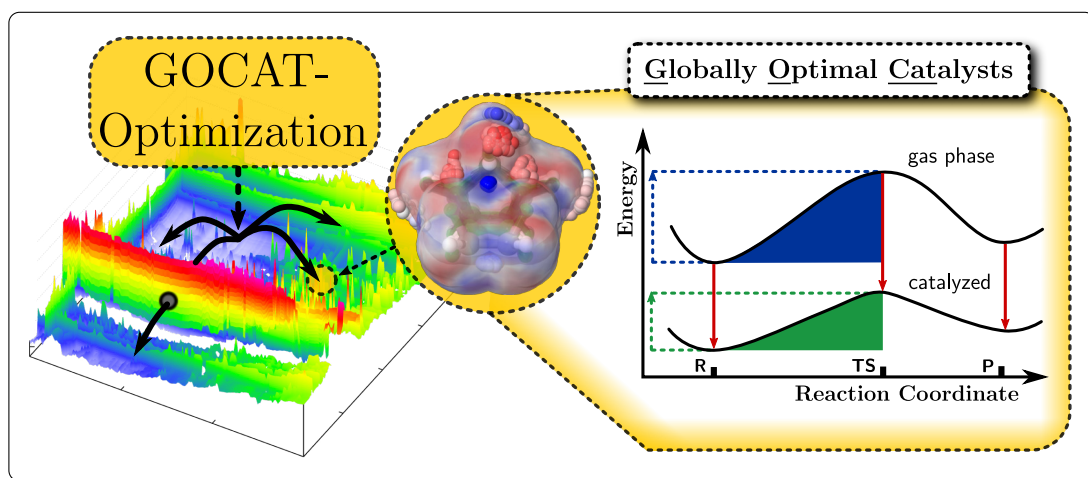


Globally Optimal Catalysts

Computational Optimization Of Abstract Catalytic
Embeddings For Arbitrary Chemical Reactions



Dissertation

in partial fulfillment of the requirements
for the degree

Doctor rerum naturalium

of the Faculty of Mathematics and Natural Sciences at
Kiel University

submitted by

Mark Dittner

Kiel, September 2019

Globally Optimal Catalysts

Computational Optimization Of Abstract Catalytic Embeddings For Arbitrary Chemical Reactions

Mark Dittner
Kiel University
Institute of Physical Chemistry
Dissertation, Kiel, September 2019

1. Referee: **Prof. Dr. Bernd Hartke**
Institute of Physical Chemistry
Kiel University

2. Referee: **Prof. Dr. Carolin König**
Institute of Physical Chemistry
Kiel University

Date of Oral Examination: September 5, 2019

Approved for Publication: September 5, 2019

Digital version, 1.0, of September 11, 2019:

Two-color scheme, vectorized graphics, no binding corrections.

Abstract

In silico design of molecules usually implies an inverse strategy, starting from desired properties and inferring molecular systems that realize them. Inverse problems, however, are hard to solve in practice. One feasible approach is to do a forward sampling by starting from many well-chosen and chemically meaningful systems and calculating the property in question directly, and then selecting the system coming closest to the desired property value. Naive realizations of this forward strategy fail because chemical space is vast even for small compounds, which makes complete enumerations and deterministic global optimization impossible. Thus, additional ingredients are necessary such as (1) imposing abstractions in chemical representation in order to further shrink and simplify the search landscape and (2) leveraging unbiased metaheuristic optimization algorithms that avoid complete enumeration. One specific composed property of chemical matter is the catalytic impact it could exert on chemical reactions. Finding such suitable systems for catalysis and understanding these poses one of the essential issues for chemists.

In this context, the long-term goal of this Thesis is to develop a general framework which tackles the design of molecular systems for an optimal catalytic effect onto arbitrary chemical reactions. For any given reaction, an arrangement of an additional molecular framework around this reaction center is sought such that the energetic reaction barrier is lowered as much as possible. Thus, for item (1), the so-called **globally optimal catalyst (GOCAT)** model is introduced, and for item (2), **evolutionary algorithms (EAs)** are harnessed as implemented in our global optimization suite for chemical problems, OGOLEM, which was highly extended to allow for these catalysis optimizations. Starting with a maximally reductionistic approach for studying the non-bonding interactions, *electrostatic GOCATs* are introduced that consist of arbitrary numbers, distributions and strengths of partial point charges around reacting molecules, mostly surrounding these on a common exposed surface. Selected proof-of-principle reactions with known catalytic trends are studied, by using different flavors of objective functions that must be defined for the catalytic enhancement. This ranges from a simple Menshutkin reaction with a clearly anticipated trend of the catalytic effect, to more subtle reactions, such as a **Diels–Alder** reaction. In the latter case, full re-optimizations of the reaction path within the catalytic surrounding are allowed, which can even result in mechanistic changes of the reaction. Moreover, some enzymatic reactions and steps in transition-metal catalytic cycles are also touched upon. The respective analysis and the understanding is facilitated by methods from inferential statistics and **machine learning**. In the current state, this framework allows to encode almost any energy and gradient property into the objective function and to optimize **GOCATs** for electrostatic interactions in order to alter the reaction profile on an effective potential energy surface. This is also in line with recent studies about pure electrostatic catalysis in both the computational and experimental realms. In fact, the general **GOCAT** model is intended to successively be improved with more complex interaction centers such as van-der-Waals centers, H-bond donors/acceptors, etc., also including bonding interactions, and, finally, to translate these abstract embeddings in a subsequent step to molecular realizations back again. These improvements, however, are the topic of further ongoing research.

Besides, many method development matters are addressed: They range from optimal shared-memory parallelization, exemplified for global parameter optimization of the reactive force field, REAXFF, *via* diversity control parameters for the **EAs**, applied to a cluster structure optimization problem, to **EA** operator benchmarks and optimizations of abstract electrostatics.

Kurzzusammenfassung

In-Silico-Design von molekularen Systemen impliziert gewöhnlich eine inverse Herangehensweise, wobei von gewünschten Eigenschaften der Moleküle zurückgeschlossen wird auf mögliche Systeme, die diese Eigenschaften aufweisen. Die Beantwortung solcher inverser Fragestellungen ist jedoch schwierig. Daher ist *ein* geeigneter Ansatz, diese unmittelbaren inversen Schritte durch Rechnungen in herkömmlicher Vorwärtsrichtung zu ersetzen, wobei von jeweilig sinnvoll gewählten Molekülen ausgegangen und anschließend das System gewählt wird, welches dem Ziel am nächsten gekommen ist. Ein zu naives Vorgehen hierbei ist der Sache jedoch abträglich, da bereits der chemische Raum potenzieller Moleküle zu extensiv ist und ein einfaches, deterministisches Durchprobieren daher das Vorhaben konterkariert. Deshalb werden weitere vereinfachende Schritte nötig: (1) Eine weitere Abstraktion der chemischen Repräsentation, um den Eigenschaftssuchraum zu verkleinern und zu vereinfachen, sowie (2) das Anwenden metaheuristischer Optimierungsalgorithmen zur unverzerrten und effizienten Suche ohne vollständiges Erkunden des Suchraums. Eine besondere Eigenschaft chemischer Systeme ist ein etwaiger katalytischer Einfluss von diesen auf chemische Reaktionen. Geeignete Katalysatoren für beliebige chemische Reaktionen zu finden und diese zu verstehen oder gar maßzuschneidern, ist ein zentrales Anliegen in der Chemie.

In diesem Kontext versucht die vorliegende Arbeit als Langzeitziel eine passende Plattform zu entwickeln, welche das generelle Design molekularer Systeme für einen optimalen Katalyseeffekt auf beliebige chemische Reaktionen projiziert. Für eine gegebene Reaktion soll eine hinzukommende chemische Umgebung komponiert werden, welche die Reaktionsenergiebarriere so weit wie möglich vermindert. Daher wird für Punkt (1) das sogenannte Modell des **globally optimal catalyst (GOCAT)** eingeführt, und für Punkt (2) kommen Evolutionäre Algorithmen (EAs) zur Anwendung, wie sie bereits in unserem Programmpaket zur Lösung allgemeiner globaler Optimierungsprobleme der Chemie, **OGOLEM**, bereitgestellt werden, welches jedoch deutlich für diese Katalysoptimierungen ergänzt wurde. Angefangen in einem maximal-reduktionistischen Ansatz werden *elektrostatische GOCATs* erarbeitet, die aus einer beliebigen Anzahl, Verteilung und Stärke von Partialladungen bestehen und rund um die reagierenden Moleküle drapiert werden, meist auf einer gemeinsamen exponierten Oberfläche. Damit werden bestimmte Grundlagenuntersuchungen bezüglich ausgewählter Systeme angestellt, deren potentielle katalytische Einflussmöglichkeit zumindest weitestgehend bekannt ist und welche durch unterschiedliche Variationen der Zielfunktionen für die maximale Reaktionsratenerhöhung untersucht werden. Diese Untersuchungen erstrecken sich von einer einfachen Menshutkin-Reaktion bis hin zu etwas subtiler beeinflussbaren Systemen wie eine Diels-Alder-Reaktion. Hier können sogar durch die Einbettung induzierte mechanistische Veränderungen verfolgt und optimiert werden, da eine vollständige Reoptimierung des Reaktionspfades ermöglicht ist. Darüber hinaus streift das Anwendungsspektrum auch enzymatische Reaktionen sowie Reaktionsschritte eines Übergangsmetallkatalysatorzyklus. Die Analyse und das Verständnis der Systeme werden dabei mittels inferentieller statistischer Verfahren sowie Maschinellen Lernens verfeinert. Der Status-Quo dieser Plattform ermöglicht bereits, beliebige Energie- sowie Gradienteneigenschaften als Zielfunktion zu kodieren, um **GOCATs** für eine maximale, elektrostatische Katalyse zu erschaffen, die das Reaktionsprofil auf der effektiven Potentialenergiefläche konvenient beeinflussen. Dies bettet sich bereits gut in aktuelle Studien sowohl theoretischer als auch experimenteller Natur ein. Tatsächlich ist intendiert, dass das generelle **GOCAT**-Modell im Verlauf der weiteren Entwicklung sukzessive auch kompliziertere Wechselwirkungssituationen mit einbeziehen kann, z.B. mittels van-der-Waals-Zentren, Wasserstoffbrücken-Donoren/-Akzeptoren und weiterer denkbaren, sowie schließlich auch bindende Interaktionen und dass final der anschließende Schritt von der Abstraktion zu echten molekularen Realisierungen ermöglicht wird. Diese Verbesserungen sind jedoch das Ziel aktueller Forschung.

Weiterhin werden unterschiedliche Methodenentwicklungsaspekte angesprochen: Diese reichen von verbesserter Parallelisierung in Mehrprozessorarchitekturen, beispielhaft gezeigt anhand einer globalen Parameteroptimierung des reaktiven Kraftfeldes **REAXFF**, über Diversitätskontrollparameter des **EAs**, illustriert mittels eines Clusterstrukturoptimierungsproblems, bis hin zu **EA**-Operator-Testevaluationen und allgemeinen abstrakten Elektrostatikoptimierungen.

Contents

Abstract	v
Kurzzusammenfassung	vii
List of Figures	xiii
List of Tables	xvii
List of Algorithms	xix
List of Acronyms	xxi
1 Introduction	1
1.1 Molecular Design in Theoretical Chemistry	2
1.2 Catalysis and its Design	4
1.3 Globally Optimal Catalysts	8
2 Theory	13
2.1 Optimization	13
2.1.1 Local Optimization	14
2.1.1.1 Penalty Function Method	16
2.1.1.2 Concrete Case of Electrostatic Potential Optimization . .	18
2.1.2 Global Optimization	21
2.1.2.1 Metaheuristic Optimization	24
2.1.2.2 Genetic Algorithm and Terminology	27
2.2 Potential Energy Surface	30
2.2.1 Hartree-Fock Approximation	31
2.2.2 Density Functional Theory	34
2.2.3 Semi-Empirical Approximation	36
2.2.4 Empirical Potential	39
2.2.4.1 ReaxFF	39
2.2.4.2 EVB-QMDFF	40
2.3 Electrostatics	42
2.3.1 Coulomb's Law	42
2.3.2 Molecule Exposed to a Non-Uniform Electric Field	43

2.4	Coupling Model	47
2.4.1	General Quantum Mechanics/Molecular Mechanics	47
2.4.2	Coupling in Purely Electrostatic Globally Optimal Catalysts	49
2.4.3	Quantum Mechanics/Molecular Mechanics with Semi-Empirical Coupling	49
2.5	Minimum Energy Path	51
2.5.1	Nudged Elastic Band	52
2.5.1.1	Improved Tangent and Climbing Image	53
2.5.1.2	Nonlinear Interpolation	56
2.5.1.3	Adaptive Nudged Elastic Band	57
2.5.1.4	Possible Improvements	59
2.6	Machine Learning	61
2.6.1	Multidimensional Scaling	64
2.6.2	Hierarchical Clustering	65
2.6.3	Distance Metric	66
3	Methodology and Implementations	69
3.1	OGOLEM	69
3.2	Overview	71
3.3	OGOLEM’s Genetic Algorithm	72
3.4	Further New Capabilities	74
3.5	More Detailed Operator Descriptions	76
3.5.1	Recombination	76
3.5.2	Mutation Operator: canada	79
3.5.3	Niching	83
3.6	Quality Assessment of Catalysis	85
3.6.1	Static Fitness Function	86
3.6.2	Faulty Fitness Function	89
3.6.3	Adaptive Fitness Function	91
4	Publication: REAXFF Parameter Optimization	95
4.1	Scope of the Project	95
4.2	Publication Data and Reprint	96
4.3	Complementary Information about Parallelization Improvements	109
5	Publication: Lennard–Jones Cluster Optimization	117
5.1	Scope of the Project	117
5.2	Publication Data and Reprint	118
6	Publication: Optimization of Globally Optimal Catalysts	127
6.1	Scope of the Project	127
6.2	Publication Data and Reprint	128

6.3	Complementary Information	147
6.3.1	Greedy Benchmark of Meta-Parameters	147
6.3.1.1	General Setting for the Benchmark	148
6.3.1.2	Results and Discussion	150
6.3.2	Transferability by Translation Protocols	157
6.3.2.1	Protocols for Translation	157
6.3.2.2	Results and Discussion	159
6.3.3	Correlation Studies	171
6.3.3.1	Primer on Electrostatic Catalysis	171
6.3.3.2	Menshutkin Reaction as Limiting Case	173
7	Adaptive Globally Optimal Catalysts	177
7.1	Overview on Recent Electrostatic Catalysis	177
7.2	Diels–Alder Reaction	180
7.3	Methodology	181
7.4	Results	185
7.4.1	Static Globally Optimal Catalysts	185
7.4.2	Adaptive Globally Optimal Catalysts	195
7.4.3	Uniform Electric Fields and Comparison	201
7.5	Discussion	206
7.5.1	Comparison with Literature	206
7.5.2	Background Statistics for the Adaptive Fitness Function	208
7.5.3	Critical View and Improvements	209
7.5.4	Conclusion	212
8	Conclusions	213
8.1	Master’s Thesis of BEHRENS	213
8.2	Summary and Conclusion	220
8.2.1	General Project Outline	220
8.2.2	Summary of Observations	222
8.3	Prospects	226
	Bibliography	235
A	Globally Optimal Catalyst Optimization	265
A.1	Operator Benchmarks for the Genetic Algorithm	265
A.2	Electric Fields for the Menshutkin Reaction	270
A.3	Clustering for the Diels–Alder Reaction and Further Data	273
A.4	Correlations for the Diels–Alder Reaction	278
A.5	Uniform Electric Field Data for the Diels–Alder Reaction	282
B	Supplementary Information for the Publications	287
B.1	ESI: REAXFF Parameter Optimization	288
B.2	ESI: Lennard–Jones Cluster Optimization	294

B.3 ESI: Optimization of Globally Optimal Catalysts	299
Acknowledgements	i
Declaration	iii

List of Figures

1.1	Generic catalytic effect	6
2.1	Illustration of the penalty terms	20
2.2	Illustration of characteristics of the objective function surface	22
2.3	Traditional genetic algorithm operators	27
2.4	Illustration of a typical inhomogeneous ESP within a GOCAT	46
2.5	PES and MEP illustration	52
2.6	Illustration of the NEB force projections	55
2.7	Illustration of interpolation schemes	57
2.8	Adaptive NEB definitions	59
2.9	Illustration of a discretized MEP generated by NEB	60
2.10	Illustration of hierarchical clustering	65
3.1	Main GA optimization cycles in OGOLEM	73
3.2	Illustration of the portugal recombination operator	77
3.3	Illustration of the sweden recombination operator	79
3.4	Problem-specific unaray search (mutation) operator: canada	81
3.5	Exemplified energies and gradient norms after canada mutation	83
3.6	Heatmaps of ESP after canada mutation	84
3.7	Simple ESP -based niching	84
3.8	Illustration of overfitting for GOCAT optimization	90
4.1	REAXFF optimization publication: graphic	96
4.2	Illustration of the explicit memory handling implementation for REAXFF optimization	112
5.1	LJ clusters publication: graphic	118
6.1	$N_{Ch} = 10$: Benchmark of GA operator settings for GOCAT design; part I . .	152
6.2	$N_{Ch} = 10$: Benchmark of GA operator settings for GOCAT design; part II .	153
6.3	$N_{Ch} = 3$ (non-neutral): fitness values after translation to DFT	160
6.4	$N_{Ch} = 3$ (non-neutral): energies and gradient norms	161
6.5	$N_{Ch} = 10$ (neutral): fitness values after translation to DFT	162
6.6	$N_{Ch} = 10$ (neutral): energies and gradient norms	163

6.7	$N_{\text{Ch}} = 10$ (neutral): GOCATs before and after translation from best rank on PM7	167
6.8	$N_{\text{Ch}} = 10$ (neutral): GOCATs before and after translation from different PM7 GOCATs to their PBE0 pendants; part I	169
6.9	$N_{\text{Ch}} = 10$ (neutral): GOCATs before and after translation from different PM7 GOCATs to their PBE0 pendants; part II	170
6.10	Electric field catalysis: limiting cases	172
6.11	Correlation plots for the Menshutkin reaction	174
7.1	DA reaction of cyclopentadiene and maleic anhydride including coordinate system definitions	180
7.2	Some features used for ESP analysis	184
7.3	$N_{\text{Ch}} = 81$ (sphere, static): reaction energy profiles for spherical GOCATs	186
7.4	$N_{\text{Ch}} = 81$ (sphere, static): illustrations of best GOCAT	187
7.5	$N_{\text{Ch}} = 81$ (sphere, static): pairwise relationships in the database	189
7.6	$N_{\text{Ch}} = 81$ (sphere, static): pairwise relationships in the best cluster	191
7.7	$N_{\text{Ch}} = 10$ (vdW , static): reaction energy profiles for the best cluster	192
7.8	$N_{\text{Ch}} = 10$ (vdW , static): illustrations of the best GOCAT	193
7.9	Electrostatic potentials and fields surrounding the TS frames of static GO-CATs catalyzing the DA reaction	194
7.10	$N_{\text{Ch}} = 81$ (sphere, adaptive): reaction energy profiles of adaptive GOCATs	196
7.11	$N_{\text{Ch}} = 81$ (sphere, adaptive): synchronous concerted mechanism	197
7.12	$N_{\text{Ch}} = 81$ (sphere, adaptive): asynchronous concerted mechanism	198
7.13	$N_{\text{Ch}} = 81$ (sphere, adaptive): asynchronous two-step mechanism (zwitterionic intermediate)	199
7.14	Electrostatic potentials and fields surrounding the TS frames of adaptive GOCATs catalyzing the DA reaction	200
7.15	Plate GOCATs (static): reaction energy profiles in the uniform electric field along z as baseline for the <i>endo</i> DA Reaction	201
7.16	Plate GOCATs (static): uniform electric field correlations	205
7.17	$N_{\text{Ch}} = 81$ (sphere, static): correlation plot for barrier decrease estimation	205
7.18	$N_{\text{Ch}} = 81$ (sphere, adaptive): correlation plots for the DA reaction	207
8.1	Selected GOCAT applications of BEHRENS; part I	215
8.2	Selected GOCAT applications of BEHRENS; part II	216
8.3	Selected GOCAT applications of BEHRENS; part III	218
A.1	$N_{\text{Ch}} = 5$: benchmark of GA operator settings for GOCAT design; part I	266
A.2	$N_{\text{Ch}} = 5$: benchmark of GA operator settings for GOCAT design; part II	267
A.3	$N_{\text{Ch}} = 20$: benchmark of GA operator settings for GOCAT design; part I	268
A.4	$N_{\text{Ch}} = 20$: benchmark of GA operator settings for GOCAT design; part II	269
A.5	Electrostatic potentials and fields for the Menshutkin reaction; part I	270
A.6	Electrostatic potentials and fields for the Menshutkin reaction; part II	271

A.7	Electrostatic potentials and fields for the Menshutkin reaction; part III . . .	272
A.8	$N_{\text{Ch}} = 81$ (sphere, static): HC for the <i>endo</i> DA reaction	273
A.9	$N_{\text{Ch}} = 10$ (vdW, static): HC for the <i>endo</i> DA reaction	274
A.10	$N_{\text{Ch}} = 81$ (sphere, static): 2D MDS	275
A.11	$N_{\text{Ch}} = 10$ (vdW, static): 2D MDS	276
A.12	$N_{\text{Ch}} = 10$ (vdW, static): all $10 \cdot 853$ charges superposed for c19	276
A.13	$N_{\text{Ch}} = 10$ (vdW, static): reaction energy profiles of the best cluster ($\Delta E_{\text{R}} > 0$ allowed)	277
A.14	$N_{\text{Ch}} = 10$ (vdW, static): all $10 \cdot 736$ charges superposed for c14 ($\Delta E_{\text{R}} > 0$ allowed)	277
A.15	$N_{\text{Ch}} = 81$ (sphere, static): pairwise relationships for the highest correlated cluster	279
A.16	$N_{\text{Ch}} = 81$ (sphere, adaptive): pairwise relationships for adaptive GOCATs .	280
A.17	$N_{\text{Ch}} = 81$ (sphere, adaptive): correlation plots for the DA reaction	281
A.18	Electrostatic potentials and fields surrounding the TS frame within a uni- form plate GOCAT	282
A.19	Uniform plate GOCATs (static, <i>endo</i>): reaction energy profiles in uniform electric fields in $\{x, y\}$ -directions (<i>endo</i> DA)	283
A.20	Uniform plate GOCATs (static, <i>exo</i>): reaction energy profiles in uniform electric fields in $\{z, y\}$ -directions (<i>exo</i> DA)	284
A.21	Plate GOCATs (adaptive): reaction energy profiles and correlation plots for the DA reaction	285

List of Tables

2.1	Overview of GA terminology and its alternatives	28
3.1	Current extent of the code base of OGOLEM	72
6.1	Selected GA benchmark settings and their descriptions	151
6.2	Descriptive statistics of fitness after translation to DFT	164
6.3	Descriptive statistics of energies and gradient norms after translation to DFT	165
7.1	Properties of the discussed GOCAT models catalyzing the <i>endo</i> DA reaction	203

List of Algorithms

2.1	Simplest gradient descent	15
2.2	Simplest hill climbing	26
2.3	General genetic algorithm	29
3.1	Typical static GOCAT fitness function	87
3.2	Adaptive NEB fitness function for GOCAT design	93

List of Acronyms

AI	Artificial Intelligence	4
ANN	Artificial Neural Network	4
ANOVA	ANalysis Of VAriance	156
API	Application Programming Interface	111
BFGS	Broyden-Fletcher-Goldfarb-Shanno	16
BO	Born–Oppenheimer	3
BOB	Bag Of Bonds	67
BSD	Berkeley Software Distribution	72
CI	Climbing Image	56
CM	Coulomb Matrix	67
CNDO	Complete Neglect of Differential Overlap	36
COSMO	COnductor-like Screening MOdel	86
CSO	Cluster Structure Optimization	26
DA	Diels–Alder	177
DFT	Density Functional Theory	3
DFTB	Density Functional Tight-Binding	38
D-LEF	Designed-Local Electric Field	179
DOF	Degree Of Freedom	51
EA	Evolutionary Algorithm	9
EC	Evolutionary Computation	2
EF	Electric Field	42
ESI	Electronic Supplementary Information	96
ESP	Electrostatic Potential	18
EVB	Empirical Valence Bond	39
FES	Free Energy Surface	52
FF	Force Field	31
FIRE	Fast Inertial Relaxation Engine	56
GA	Genetic Algorithm	2
GdMC	Gradient-driven Molecule Construction	8
GFN-xTB	Geometry, Frequency, Noncovalent, eXtended Tight Binding	38
GGA	Generalized Gradient Approximation	36
GNU GPL	GNU General Public License	72
GOCAT	Globally Optimal CATalyst	9
GPR	Gaussian Process Regression	61
HC	Hierarchical Clustering	65
HDNNP	High-Dimensional Neural Network Potential	63
HF	Hartree–Fock	32
HPC	High Performance Computing	70

IDPP	Image Dependent Pair Potential	56
IEF	Interfacial Electric Field	179
INDO	Intermediate Neglect of Differential Overlap	36
IRC	Intrinsic Reaction Coordinate	51
JNI	Java Native Interface	111
JVM	Java Virtual Machine	111
KS	KOHN-SHAM	35
LBFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno	16
LCAO	Linear Combination of Atomic Orbitals	33
LCAP	Linear Combination of Atomic Potentials	3
LDA	Local Density Approximation	36
LEF	Local Electric Field	178
LJ	Lennard-Jones	9
MC	Monte Carlo	2
MD	Molecular Dynamics	47
MDS	MultiDimensional Scaling	64
MEP	Minimum Energy Path	5
ML	Machine Learning	4
MM	Molecular Mechanics	10
MNDO	Modified Neglect of Differential Overlap	37
MO	Molecular Orbital	32
MPP	Massively Parallel Processing	73
NDDO	Neglect of Differential Diatomic Overlap	36
NEB	Nudged Elastic Band	53
NFL	No Free Lunch	147
NN	Nearest Neighbor	185
OCF	Optimal Catalytic Field	5
OEEF	Oriented External Electric Field	177
OO	Object Orientation	109
P	Product	5
PBE	Perdew-Burke-Ernzerhof	36
PCA	Principal Component Analysis	64
PD	Potential Derived	19
PES	Potential Energy Surface	5
PM7	Parametrized Model 7	36
QM	Quantum Mechanics	1
QMDF	Quantum Mechanically Derived Force Field	39
QSAR	Quantitative Structure-Activity Relationship	67
QSPR	Quantitative Structure-Property Relationship	67
R	Reactant	5
RESP	REstrained Electrostatic Potential	20
RMI	Remote Method Invocation	70
RMSD	Root-Mean-Square Deviation of Atomic Positions	66
SAS	Solvent Accessible Surface	20
SCF	Self-Consistent Field	32
SD	Slater Determinant	31
SLOC	Source Lines Of Code	71
SMACOF	Scaling by MAjorizing a COmplicated Function	64

SMP	Symmetric MultiProcessing	73
SOAP	Smooth Overlap of Atomic Positions	67
SP	Single Point	33
SQC	Semi-empirical Quantum Chemistry	33
STM	Scanning Tunneling Microscope	178
TISE	Time-Independent Schrödinger Equation	30
TS	Transition State	5
TSS	Transition State Stabilization	5
TST	Transition State Theory	5
UPGMA	Unweighted Pair Group Method with Arithmetic mean	65
vdW	van der Waals	7
ZDO	Zero Differential Overlap	36
ZPE	Zero Point Energy	39

Introduction

Finding chemical systems with tailored properties is one of the essential external demands on chemistry, as well as one of the internal claims arising within it. This endeavor presupposes a fundamental understanding of relations between the chemical structure of the system and its emergent properties or “function” in certain contexts. Consequently, this understanding allows prudential chemists to “tweak” or modify such systems *rationally* in order to converge to the intended goal, which is why it is also called “designing”.^[1] In general, this holds true for all disciplines in chemistry and, naturally, for all related fields using such deductive and reductionistic reasoning. Over the decades, important discoveries, however, often included venturesome trial-and-error procedures as well as a great portion of mere luck. Common examples for such serendipities, also exceeding chemistry, are the discoveries of X-rays, cosmic microwave background radiation, penicillin, super glue and Teflon, to mention a few.

The manifest question behind such undertaking recast from the perspective of the 21st century is the role and its extent computational sciences can play here.^[2] In fact, the laws governing a large part of physics and the whole chemistry are known for almost 100 years, as DIRAC once famously pointed out.^[3] This has led to the main program of how to apply these laws and to numerically solve the resulting equations. Given a chemical structure, indeed, this can be input to the Schrödinger equation describing the matter on that scale where (non-relativistic) **quantum mechanics (QM)** is necessary which then leads to exact predicted properties, at least in principle. This can be called the common *direct* or *forward* approach. Only the smallest systems can be solved analytically in practice, though. Consequently, meaningful numerical approximations for real-life problems are crucial, manifold and intensively researched, which leads to the question: How can theoretical and computational chemistry address this theme of designing chemical matter *in silico* driven by all the capabilities of contemporary hardware and algorithm progression?^[2]

When the agenda is reformulated in a more precise form from the outset, what is really needed here is a transition from a target property to a chemical structure and composition, *i.e.*, in a direction *inverse* to the usual domain of **QM** descriptions.^[4–8] The inversion of the corresponding differential equations in a clear *mathematical sense*^[9] has been shown to

be possible only for very simplified (one-dimensional) model systems,^[10] where simple generalizations are ill-posed, in principle.^[11] This historical line is very clearly drawn in Ref. [7]. However, more research is definitely needed, which is also pointed out in a more optimistic manner recently.^[8] Additionally, inverse transitions from experimental data to the underlying Hamiltonians encoding the matter can also have some beneficial properties,^[12] which is also true, *e.g.*, when experimental parameters are optimized directly.^[13,14] Hence, what are the possibilities nowadays for tackling such inverse problems?

1.1 Molecular Design in Theoretical Chemistry

Generally, the number of just the small organic and chemically feasible compounds is estimated to be on the order of 10^{60} .^[15-17] Obviously, this chemical space^[18] is vastly beyond any imagination, and it is also palpably impossible to synthesize all these molecules or to computationally predict each system's properties *exhaustively*. For instance, one of the largest chemical collections in a virtual database made so far, the chemical space project,^[19] catalogs about 166.4 billion organic molecules up to 17 atoms each, and a subset of their properties have already been predicted.^[20,21] Hence, this extent of the total chemical space might seem intimidating, but it actually portrays an avenue for further search in view of what has *not yet* been explored so far. In accepting this challenge, current research has developed a spectrum of methods for an *efficient* and somehow *guided* search through this space without having to tackle each compound thoroughly or, again, simply randomly.

One of the traditional efforts can be traced back to high-throughput virtual screening,^[22] which started mainly in the context of pharmaceutical drug discovery (*e.g.*, see Ref. [23]). As an explorative tool, based on pre-defined or generated databases, promising molecules with targeted properties are searched for that can be investigated further in subsequent experimental or more detailed studies. Arguably, this might seem as an ensemble version of the direct or forward approach,^[24] but the “philosophy” of this data-driven, automated and heuristical methodology is not very different from methods that will be mentioned next.^[22]

When an objective as, *e.g.*, a difference between a sample's property and the intended one is formulated, this design problem can be mapped to an optimization problem of the parameters that describe the chemical matter. Clearly, as the mathematical direct inversion is out of reach, guided search procedures leveraging all kinds of different search and optimization algorithms are vital to somehow efficiently sample the promising regions and *not* all the search space, that is, chemical space. These techniques reach from greedy versions such as gradient-based optimizations, to randomized (so-called metaheuristical)^[25,26] approaches, including general **monte carlo (MC)** methods,^[27] specialized versions thereof such as simulated annealing,^[28] and general **evolutionary computation (EC)**^[29] which subsumes the **genetic algorithm (GA)**,^[30,31] evolution strategies, particle swarm optimization, besides many more. General designs for ample properties can be found in the pertinent reviews and perspectives (and references therein).^[5-8,24,32-34] In the following, some of the more common techniques that have been developed for inverse design will be described briefly, before the actual theme of this Thesis is introduced in the next Section.

Other research on molecular design: Being able to leverage gradients of differentiable smooth functions in this design problem-class is clearly advantageous for having an informed best guess about the local search direction. This also motivated the introduction of specific continuous representations of the design problem, *e.g.*, in the form of the so-called **linear combination of atomic potentials (LCAP)**^[35] as well as alchemical potentials^[36,37] for transformations through chemical space. Besides the universal kinetic and electron repulsion parts, from the **Born–Oppenheimer (BO)** separated electronic Hamiltonian follows that the system-specificity is based on both the total electron number and the Coulomb attraction potential between electrons and nuclei only, which is also called external potential in the **density functional theory (DFT)** perspective.^[38] This external potential encodes the nuclei positions and their atom types that are present and can be subject to optimization in order to reach the intended properties. By changing the potential, the chemical system is transformed into another one, thus leading to moves through chemical space, though usually by fixing the spatial positions of the atoms.

Since not each such transformed potential maps back to a realizable molecule, WANG *et al.* proposed the idea to expand such an external potential in a linear combination of (pre-defined) atomic (or atomic group) potential functions, the **LCAP**.^[35] The linear expansion coefficients varying between zero (absence) and one (presence) can then be optimized instead, and by simply rounding the highest coefficients at the end, representability is enforced. However, local minima can trap such an optimization. Hence, other movesets were introduced in that same framework that can also jump between the representable points in search space directly.^[39–41] This scheme has already been used in rather sophisticated designs such as, *e.g.*, optimizing the first hyperpolarizability of porphyrin materials,^[42] optimizing the acidity of 2-naphthols^[43] and in protein design.^[44,45] In a tight-binding extension,^[46] optimal photoabsorbers for dye-sensitized solar cells were optimized^[47] as well as Ni(II)-based catalysts for CO/CO₂ conversion.^[48] In the latter case, the ligand composition was optimized to reduce the activation energy of the rate-limiting step in a catalytic cycle.

In much the same fashion, the concept of nuclear chemical potentials haven been developed by VON LILIENFELD *et al.*^[36,37] This nuclear chemical potential is the derivative of the total electronic energy with respect to the nuclear charge distribution. By changing the nuclear charge distribution, the chemical system is transformed. Such alchemical transformations are based on the concept of state functions which lead to the fact that the actual path taken between two such states is not important. Experimentally (or chemically) unrealistic state changes made *in silico* are thus called alchemical. Similar schemes of a continuous transition between two states (or model Hamiltonians) are known from thermodynamic integration^[49] and variations thereof. Generally, this is also related to the **LCAP** approach above. Within a rigorously defined grand-canonical **DFT** framework,^[37] a coupling parameter between the states can be used to induce these alchemical transformations. Here, one molecule can be continuously changed to another usually iso-electronical molecule by interpolating between their external potentials. In first applications, interaction energies,^[50] or the doping of benzene rings with other elements were investigated.^[51] Ana-

lytical derivatives between any pair of isoelectronic systems were also developed.^[52] These derivatives can be used to guide the search through chemical space or to estimate a large number of energies (for alchemical neighbors) that are based on only a few calculations. This, too, was also extended to estimate energy barriers of simple reactions directly.^[53] In a similar vein, in a recent application the binding energies for adsorption of oxygen species on alchemically transformed slabs of different alloys (Pt, Pd, Ni) were tackled.^[54] Again, it was emphasized that after the accuracy of the alchemical derivatives of given systems has been assessed, the actual screening for better materials is straightforward with only a few **DFT** calculations. Note that there are many more studies using this general framework also in other work groups. To get an overview of these studies, the interested reader may be referred to a recent review.^[34]

Another promising approach to this problem of chemical design that recently enjoys great popularity is the utilization of **machine learning (ML)** techniques, which is itself a sub-field of **artificial intelligence (AI)** (and will be partially reviewed later in this Thesis in Section 2.6). As a data-centric approach, structure within and from data that is not needed to be evident is inferred which can lead to a plethora of varying techniques and possible applications. Only two examples are mentioned here with the focus to either introduce performative models for predicting molecular properties, in *e.g.*, Refs. [55–58] (bypassing the Schrödinger equation) or to build models to *generate* chemical systems.^[24,33,59] The latter includes a compression in feature space (or latent space) by respective deep **artificial neural network (ANN)** architectures to learn representations of chemical matter—that, again, enables to interpolate in chemical space—and prediction layers under reinforcement learning to reach the intended properties.

On the whole, when taking a step back and looking at the big picture—although we have barely scratched the surface yet—which contribution is *this* Thesis going to make in all this?

1.2 Catalysis and its Design

Catalysis is literally indispensable to life, which is only possible by all the little kinetic helpers in all types of intricate chemical networks based on the most powerful naturally evolved catalysts, the enzymes, that keep the metabolic processes going. Yet, exceeding this metaphorical view, kinetic control of chemical reactions *per se* is one of the tantalizing challenges in chemistry.^[60] A catalyst is formally defined to be a substance that is not consumed during a chemical reaction and can increase the reaction rate of specific reactions extremely.^[61] This can make the difference, for instance, for biochemical processes (or metabolism) occurring under ambient conditions at a sufficient pace at all. This fact can also be a source of inspiration for a variety of applications.^[62] Furthermore, also up to 80% of manufacturing processes involve somehow catalytic steps for proficient productions.^[63] As one of the central topics in chemistry, catalysis has been extensively studied ever since.^[64] Yet computer-aided designs are still very challenging.^[60,65]

Thus, the target property that is taken up in this Thesis is catalysis in general. This can be

thought of as a composed energetic property on the **potential energy surface (PES)** for the molecular system where the corresponding activation energy is supposed to be minimized. Fig. 1.1 on the following page illustrates this in terms of a generic energy surface. Generally, the **reactant (R)** and **product (P)** of certain reactions must be identified as (local) minima on the **PES**, and the corresponding mechanism can be assigned to a **minimum energy path (MEP)** that connects these minima.¹ When neither multiple steps with many other local intermediates as consecutive reactions nor many more parallel reactions along different paths occur, the only lowest-energy first-order saddle point, the **transition state (TS)**, is pivotal for the (statistical) kinetic depiction. In this vein, the usual, well-known **transition state theory (TST)**^[66] often suffices for the description of these thermal reactions:^[67-69]

$$k(T) = \kappa \frac{k_B T}{h} \exp\left(-\frac{\Delta G^\ddagger}{RT}\right), \quad (1.1)$$

as exponential relation between the chemical rate k at a temperature T with the Gibbs free energy barrier ΔG^\ddagger between **TS** and **R**. (Note that this Eq. (1.1) is the unimolecular variant; otherwise standard state references as well as a concentration factor for correct dimensions would be present.) k_B is the Boltzmann, h the Planck, R the universal gas constant and κ the transmission coefficient that essentially corrects for some of the underlying assumptions made in **TST** (e.g., dynamical recrossings, **QM** tunneling).^[70] In Eq. (1.1), the Gibbs free energies may be approximated by the dominating electronic energies at zero Kelvin first of all, excluding entropy and temperature effects in this way. No absolute rates will be calculated anywhere in this work, instead Eq. (1.1) shall just recall the proportionality between such a rate and the energetic barrier, given as ΔE^\ddagger in Fig. 1.1 on the next page, that is to be minimized for a catalytic effect.

Before the actual new approach of this work is clarified, some concepts for catalysis are recapitulated as well as similar approaches are contrasted, first of all. In the following, the focus will lie on rather well-understood systems for homogeneous catalysis. The advantage in treating homogeneous catalysis computationally is simply the facilitated comprehensibility compared to all the complexities that come with the modeling of heterogeneous catalysis on surfaces.^[71,72]

Other research on catalysis: Reaching back to one of the first proposals of how enzymes work, PAULING^[73,74] has stated that proficient catalysis is based on the complementary shape of such enzymes to the **TS**, which was formalized later by WOLFENDEN.^[75] In this regard, the focus was placed on **transition state stabilization (TSS)** relative to the substrate for the reaction rate increase (see also the historical line in Ref. [76]). Further energetic decomposition schemes as well as optimal (electrostatic) surroundings were introduced by SOKALSKI,^[77-79] whose ideas are best explained in Ref. [80]: The optimal surroundings that are called **optimal catalytic fields (OCFs)** are essentially computed by a molecular electrostatic potential calculation on a common surface of both the superposed substrate

¹ Note that the acronyms “R” and “P” are used consistently throughout this Thesis for denoting both the singular and the plural forms, reactants and products, if needed.

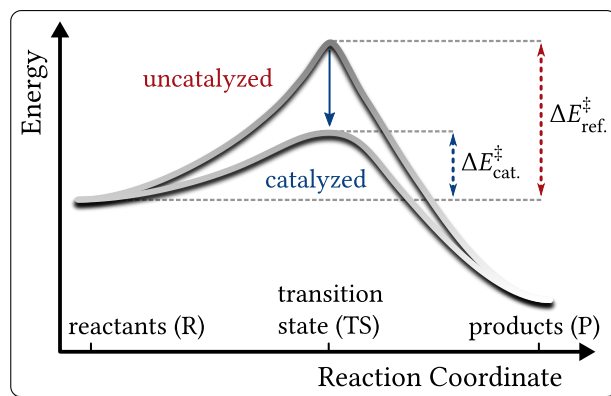


Fig. 1.1: Generic reaction energy barrier decrease by the presence of a catalyst. Note that one could argue whether the reaction coordinate is the same for both paths, whether no stabilization effects occur at **R** or **P**, whether rather Gibbs free energies should be shown instead, or other intricacies that are not insinuated by this simplified illustration.

(or **R**) and **TS** structures of the reacting molecule(s). The difference between the two of them is then evaluated and the sign inverted for a complementary negative image in order to assess the energetic relational change of the **TS** energy vs. the **R** energy when *one* probe charge would be placed onto the surface. Furthermore, the corresponding gradient field of an **OCF** was developed as electrostatic field for minor energetic changes with respect to the charge embedding around the **TS**. Since the relation of **R** and **TS** was addressed directly, it was called *differential TSS*. Naturally, this simple **OCF** concept is only useful for cases where electrostatic stabilization effects are dominating, which is validated by their energy decomposition scheme.^[81] The energy contributions are usually estimated beforehand, when each residue role (under the premise of two-body additivity) is investigated. These ideas have been applied multiple times on different complexity levels since then, including simple reactions and also full enzymes.^[80,82-91] With an **OCF**, simple electrostatic barrier lowering analyses and predictions can be made. This description, however, only covers the placing of *one* charge around two *fixed* structures, the substrate and the **TS**. Thus, neither synergistic polarization of the internal structures nor other information of the **PES** is incorporated. Hence, the **OCF** prediction can also become contradictory to the observed residuals within an enzyme.^[88]

To the current author's knowledge, this work of SOKALSKI *et al.* is one of the first appearances of *abstract* electrostatic environments for catalysis so far. Similar electrostatic potential based concepts were also used by others.^[92-95]

Dwelling still upon the subject of catalysis design for enzymes for the moment, rather *concrete* interactions by explicit interaction groups were developed by HOUK *et al.* in the theozyme concept.^[96-100] Here, model amino acid residues are (usually) locally optimized as a yet artificial surrounding to stabilize the **TS** structure of a reaction. These theozymes can subsequently be the central unit for a full-blown protein design. In a simplified picture, side-chains of the theozyme are linked to a protein backbone that is also supposed to fold in a way such that the pre-optimized theozyme actually is conserved, which was first addressed by MAYO^[101,102] and BAKER.^[103] Evidently, such very specific protein designs are

not the topic of this Thesis and thus no further details are given here.^[104,105] In this regard, problem-specific knowledge in form of pre-generated libraries of, e.g., good backbone snippets is again an essential part and also in form of specific energy functions.^[106] Despite impressive results of such *de novo* designed enzymes,^[107] it is also pointed out that weak binding to the substrate often occurs^[108] since the whole optimization is based on the **TS**-optimal complementary theozyme alone.^[109] This can even result in a too strong binding of the **TS**.^[110] As a corollary, some better balanced descriptions of the whole reaction path might be beneficial. Besides, some “translation error” is certainly expected to come to pass when simple (more artificial) models are translated or build-up to real functioning molecules, *i.e.*, proteins in this case. Such final enzyme designs therefore include already an experimental post-optimization procedure to better “adapt” to reality.^[107,111]

This leads to another main issue that is still unresolved until this day, namely the way enzymes actually work mechanistically. One proposition is most vigorously represented by WARSHEL, who has already emphasized the electrostatic origin as the main stabilizing effect in enzymes many decades ago,^[112,113] and who advocates this regularly since then.^[114–117] Here, the enzyme pocket is supposed to create a specific environment complementary to the **TS** with dominating electrostatic interactions, which is also called preorganization of the enzyme. This is in contrast to other works that frequently propose some small- to large-scale motions of the enzymes to play a pivotal role, which can be subsumed under the dynamics proposal (e.g., Refs. [69, 118, 119]). As long as this (outer) motion of the enzyme is just correlated with the inner reaction, this simply *is* the reaction coordinate or belongs to it and such “dynamics” are to be excepted, WARSHEL argues,^[116,117] which also includes pre-stabilized structures of the inner molecules that would not be stable in the gas or solvent phase without an enzyme. Thermal **TST** would still suffice for the description (maybe with small adjustments of the pre-factors) since it is not observed that energy is directly shuffled into specific modes as non-equilibrium effect. Often, such assumed “dynamics” is a misunderstanding when simply correlated motion, etc., is actually meant.^[116] This is also in line with recent theoretical and experimental research (that will be reviewed later in this Thesis in Sections 6.3.3.1 and 7.1).^[120–128]

Although some arguments *for* this electrostatic proposal have been put forward now, no specific side is taken in the present work since enzymes and all the resulting open questions are completely outside the scope of this Thesis. Yet, these matters point to the fact that electrostatics is a meaningful first starting point for the model presented below, in any case.

To stress this fact, Coulomb interactions (including polarization) between all kinds of seemingly different interaction proposals were argued to be the only basis of real non-bonding interactions.^[129–132] Common explanations of **van der Waals (vdW)** interactions, charge transfers, σ -hole interactions, π -stacking, and others, are either just an unreal manifestation of the mathematical (approximated) form or can be traced back solely to the electrostatic (Coulomb) interactions—which is not very surprising bearing in mind that the **QM** Hamiltonian, besides the kinetic part, only includes exactly the Coulomb terms. However, these authors of Refs. [129–132] do not dare (yet) to also include bonding interactions into this simple classical picture,^[133] which might neglect the **QM** nature of

the latter induced by the exchange and quantum correlation effects between the electrons. Thus, this is again generally a strong argument in favor of possible electrostatic impacts on molecules and thus also for catalysis.

Finally, one additional catalytic *ansatz* must be mentioned that bears the strongest resemblance to the current work. WEYMUTH and REIHER proposed the **gradient-driven molecule construction (GDMC)** approach in Refs. [7, 134]. The goal is to stabilize a molecular pre-defined fragment with desired properties by an arbitrary surrounding. Then, both the inner fragment, which is, for example, usually taken to be a proper but not-yet stationary **TS** structure as the activated fragment, as well as the new surrounding that is to be found, shall show a zero gradient with respect to the nuclear coordinates. This criterion was also generally developed in form of a so-called jacket potential equation (again as an external potential similar to the approaches above) that would lead to mutually perfectly compensated gradients. Since solving this problem would lead to the concurrent search for the optimal electron number, nuclei number, their positions and charges (*i.e.*, atom types) while all component-wise gradients of the fragment and embedding should become zero, this was (usually) split into a two-step procedure. First, an optimal jacket potential was to be found and, second, a representation of that same jacket potential. The following examples were investigated:^[134] (1) the optimizations of potential values directly on the **DFT** exchange–correlation integration grid, (2) the optimization of partial point charges either anywhere in space or at discrete space points (*i.e.*, at meaningful ligand positions for the transition-metal inner fragment) and (3) the optimization of shell-wise concrete (saturated) atomistic fragments as ligand spheres. This concrete optimization of real atomic ligands was also extended in Ref. [135].

To conclude, there are simplistic electrostatic models already available (*cf.* **OCF**), concrete saturated atomistic embeddings for enzymes (*cf.* the theozyme concept) or for other systems (transition-metal complexes). All of them explicitly or implicitly rest upon **TSS** with a focus on a proper pre-defined **TS** structure—while actually enforcing gradients to be zero can lead to *any* stationary point on the **PES** (*cf.* **GDMC**). Additionally, the aforementioned most efficient global optimization algorithms for proper short-cuts in the inverse design problem are not leveraged when having to treat again a simple model calculation (*cf.* **OCF**), using local optimizations (*cf.* theozymes), or greedy sequence-based optimizations (*cf.* **GDMC**).² This *can* bring in some unforeseen bias in principle. The concentration on only pre-defined **TS** structures should be lifted to include more information about the reaction path (*i.e.*, also the *differential TSS*) as well as further ingredients for the objective functions that are discussed later. This finally leads to the approach taken in this Thesis.

1.3 Globally Optimal Catalysts

The aim of this Thesis is the development of a framework for general optimal catalytic embeddings for arbitrary chemical reactions. In the following, they will be referred to as

² Indeed, the current author is aware of the use of differential evolution in the most abstract optimizations in Ref. [134]; but we think that improvements in this regard can be made.

globally optimal catalysts (GOCATs). Similar to all the aforementioned approaches, the GOCAT design tackles first and foremost the optimal stabilization of the respective TS structures by using the most efficient unbiased metaheuristic optimization algorithms, especially variants of evolutionary algorithms (EAs). With these, the efficient *guidance* on the property surface, *i.e.*, multiple separate objectives leading to catalysis, is incorporated for the inverse design. In order to be able to shrink the chemical compound space sufficiently, however, an additional abstraction layer must be involved. As discussed above, direct optimization within the whole chemical compound space is not feasible such that either (smaller) pre-defined libraries have been used (as in the screenings, LCAP, ML-techniques, theozymes, also GdMC in the concrete atomistic case) or such that an *abstraction layer* has been included for search space reduction essentially.

In the philosophical view of reductionism (or in a bottom-up approach),³ the GOCAT framework starts first with supposedly the simplest models possible for the catalysis and, subsequently, incorporates more concrete information and lifted model restrictions at later stages. Thus, the basis for catalytic effects can be understood and re-modeled first, while still including many (also very strict) model restrictions that might seem far away from the usually very complicated details of concrete real-life catalysts. After establishing such a basis, the model is successively improved to enclose more and more (real) chemistry. In these later stages, the GOCAT can embody essentially all the aforementioned approaches of Section 1.2 and is not supposed to be restricted to anything specific such as enzymes, transition-metal complexes, etc.

Hence, the question arises on how to start with such an agenda? As motivated above, the simplest and palpably most important embedding in *abstract* (non-atomistic) form is the electrostatic interaction that can be modeled as classical Coulomb interaction as a first approximation by, *e.g.*, point charges (similar to the first GdMC models). More complex abstract interaction entities could then be imagined such as vdW interactions, H-bond acceptors or donors, bonding type interactions, maybe even (mechanical) forces (in the realm of mechanochemistry).^[138] Naturally, these interactions must be somehow represented around the chemical systems. For instance, vdW interactions could be modeled *via* simple Lennard-Jones (LJ) potentials or by rare-gas atoms and pseudo-potentials, H-bond centers by suitable combinations of partial charges and vdW centers or with explicit molecules (such as H₂O). Bonding interactions are naturally modeled then by (un-)saturated interaction partners of a (pre-defined) library, which would lead to a concurrent discrete as well as continuous optimization problem. However, as little disclaimer not to take this the wrong way, this is not meant as a secluded list but rather as a source that is to be amended for possible representations and future models. Moreover, many of the actually already anticipated improvements have *not* yet been implemented during this Thesis, and this is the subject of further current research. These possible extensions will become apparent in the course of the present work and will be discussed at many places. At the end, the bias

³ One could also start holistically as, *e.g.*, in Ref. [136] and optimize macroscopic reaction conditions directly. This works also well for the heterogeneous regime by optimizing complete microkinetic models of reaction networks to reach intended macroscopic outputs without starting at a detailed atomistic understanding of each step.^[137]

by the model abstractions and further meta-parameters that is always still incorporated is to be gradually lifted, at least for the concrete problems at hand. To claim, however, to be able to solve all such catalyst design problems would be quite presumptuous; hence, this framework is intended to be *one* contribution to the arsenal of chemists to understand their catalytic systems and to predict better ones based on their current problems.

The real content of the **GOCAT** models that are used will make up a great portion of this Thesis (*cf.* Chapters 2 and 3). Besides, some method-developments that seemingly have nothing to do with **GOCAT** design are discussed along the way and the final proof-of-concept designs which begin with selected known reactions and model settings start later in Chapter 6 on p. 127 and on all the following pages. At these places, the concrete details and restrictions of the models will always be critically discussed.

Idea of a simple electrostatic GOCAT: Hence, before coming to improved models, the reader might think of an abstract **GOCAT** as used at many places to be constituted by the following ingredients:

- First of all, a *static/fixed* pre-optimized reaction path is used *via* a discretized **MEP** into frames (*cf.* Section 2.5) for a chemical reaction at hand. Later mechanistic changes are included by automatic full relaxation protocols.
- All frames are maximally *compact* or aligned for the reaction path.
- *One* **GOCAT** surrounds all these frames at once.
- **GOCATs** are build of N_{Ch} partial point charges that carry q_i as charge value and that sit at \mathbf{r}_i around the reaction path.
- The restricted Cartesian space is usually a specific curved 2D surface such as a *common* **vdW** surface exposed by all atoms of all frames.
- Usually, partial charges with $q_i \in [-1.0, +1.0]$, a minimum distance as $r_{ij} \geq r_{\text{min}} = 1.0 \text{ \AA}$ between two charges, i and j , and a constant total summed charge as, *e.g.*, $\sum_i^{N_{\text{Ch}}} q_i = \text{const.} = 0$ (neutrality) are to be conserved.
- **QM/molecular mechanics (MM)** as a coupling method between the outer **GOCAT** and the inner reaction is used (*cf.* Section 2.4).

The objective function will mostly include differential **TSS**, simplification penalties and also **GdMC** (in a loose form). All these model parameters are completely open for modifications for the respective problems, but—as it will become apparent—these aforementioned ones are the most meaningful ingredients to study first electrostatic unbiased catalysis optimizations.

With electrostatic **GOCATs**, already a great portion of the catalytic influences can be studied and explained. In fact, there are many recent theoretical and experimental studies that even treat sole electrostatic catalysis as an end in itself (*cf.* Sections 6.3.3 and 7.1). In using this two-step procedure of first optimizing an abstract embedding, the hard subsequent task then is to translate these back to real molecules again. Therefore, the approaches of

LCAP, of alchemical potentials and of generative models were addressed. These translation steps, however, are outside the scope of this Thesis, and they would need more extensions of the **GOCAT** design framework in the future as well as more notable research catalysis questions than the first proof-of-concept studies pursued in the present work.

Theory

This Chapter gives an overview of the important concepts that are used in this Thesis. First, general optimization is introduced in Section 2.1, both for local and global search. Secondly, Section 2.2 briefly outlines the selection of energy descriptions of chemical matter used in this work. Thirdly, depictions of general electrostatics, its specific influence on molecules and the coupling model for its description is discussed in Sections 2.3 and 2.4. Afterwards, reaction path optimization as used for later Sections is introduced in Section 2.5, and, finally, Section 2.6 summarizes some background information on the analysis methods used in later Sections, which touches on the domain of **machine learning**.

2.1 Optimization

Without any loss of generality, the explanations can be limited to the minimization problem in the following since a maximization problem can be changed to the former by a simple sign inversion of the objective function, $f(\mathbf{x})$, that is to be minimized. A mathematically concise definition of optimization can then be given as^[139,140]

$$\min_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x}) \quad (2.1)$$

$$\text{such that } g_i(\mathbf{x}) \leq 0, \quad i \in I = \{1, 2, \dots, m_1\}, \quad (2.2)$$

$$h_j(\mathbf{x}) = 0, \quad j \in E = \{1, 2, \dots, m_2\}. \quad (2.3)$$

Here, $\mathbf{x} \in \mathbb{D} \subseteq \mathbb{R}^n$ is a vector of elements x_1, x_2, \dots, x_n such that the objective function, $f(\cdot)$, is as small as possible while the m_1 inequality, Eq. (2.2), and m_2 equality, Eq. (2.3), constraints given *via* the functions $g_i(\cdot)$ and $h_j(\cdot)$ are true, respectively. All functions here are real-valued ones $f(\mathbf{x}) : \mathbb{D} \mapsto \mathbb{R}$, $g_i(\mathbf{x}) : \mathbb{D} \mapsto \mathbb{R}$, $h_j(\mathbf{x}) : \mathbb{D} \mapsto \mathbb{R}$. This is called *single-objective* optimization. Instead, if $\mathbf{f}(\mathbf{x}) : \mathbb{D} \mapsto \mathbb{R}^k$ with $\mathbf{f} = \{f_i : \mathbb{D} \mapsto \mathbb{R} : i \in 1 \dots k\}$ were used, this would be called *multi-objective* optimization.^[29] Only single objectives will be targeted in this work.¹ If Eqs. (2.2) and (2.3) hold, X defines a set of feasible points for

¹ This is less restrictive than it may seem at first: In practice, *multiple* (in-)dependent objectives are used at the same time in this Thesis. However, a strict protocol for comparing and merging the objectives into a

\mathbf{x} and it is defined as the set $X = \{ \mathbf{x} \mid g_i(\mathbf{x}) \leq 0, i \in I; h_j(\mathbf{x}) = 0, j \in E \}$. It can be called *search space* and its elements *candidate solutions*. There are different subtypes of such optimization problems with *linear* or *non-linear* constraining functions as well as objective functions, but we will stick to the most general (and practical) case here, *i.e.*, all functions are supposed to be *non-linearly* dependent on \mathbf{x} (which is also called *non-linear programming* in mathematics). If further $\mathbf{x}^* \in X$ and if

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \forall \mathbf{x} \in X, \quad (2.4)$$

then \mathbf{x}^* is the point of the *global minimum* of the problem of Eqs. (2.1) to (2.3) (also named the *global minimizer*² in mathematics). There is no other point in the feasible set that maps to a lower (or equal) value of the objective function than \mathbf{x}^* , the *optimal solution*.

If $\mathbf{x}^* \in X$ and there is a neighborhood $B(\mathbf{x}^*, \delta)$ around \mathbf{x}^* , a *local optimum* can be defined by

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \forall \mathbf{x} \in X \cap B(\mathbf{x}^*, \delta), \quad (2.5)$$

where $\delta > 0$ describes a finite region around \mathbf{x}^* for the set $B(\mathbf{x}^*, \delta) = \{ \mathbf{x} \mid \|\mathbf{x} - \mathbf{x}^*\|_2 \leq \delta \}$. This means, in the neighborhood of a distinct point \mathbf{x}^* , all function values are greater (or equal) compared to the value right at \mathbf{x}^* .³

Thus, a local optimum is at least as good as nearby elements within the neighborhood, whereas the global optimum is at least as good as any candidate solution of the feasible set. If the problem is *non-convex*, there usually exist several local optima of different quality separated by a “barrier”, *e.g.*, a concave region. In general, the objective function as well as the constraints can also be non-continuous or non-differentiable.⁴

2.1.1 Local Optimization

In the following, the goal is a local non-linear optimization of a function without constraints. The simplest well-known, yet fundamental first-order algorithm, the gradient descent, is given in Algorithm 2.1 on the next page. The Algorithm illustrates a minimization problem. Supposing a continuously differentiable function $f(\mathbf{x})$, the direction of *steepest* descent is precisely along the negative gradient, $-\nabla f(\mathbf{x})$, of that function. There are different methods available to set the step length α . Also an *exact* line search can be done in order to find an α step such that the minimum in the sub-problem along the 1D direction of the negative gradient is *exactly* found. In this case, this *ensures* convergence, generally, but usually a very slow one because subsequent iterations will partially destroy or undo the progress of

single (real) number is utilized, as it is very often the case. This is simply done as a *linear aggregation* of single objectives: $f_{\text{sum}}(\mathbf{x}) = \sum_i^n w_i f_i(\mathbf{x})$ with weights $\{ w_i \}$ which have to be specified. In principle, with a multi-objective optimization setting, this metric comparison by defining weights is simply *not* done *yet* and delegated to the final solutions and its analysis.

² Note, though, that we will not follow this particular terminology in this Thesis as there might be confusion about whether the point or the algorithm/solver is named in this way.

³ For this local and global optimum definition, there are also *strict* versions with “>” instead of “≥”.

⁴ Using an objective function with many qualitatively different ingredients can also lead to all types of errors. In the domain of the Thesis, errors occur simply due to, *e.g.*, convergence issues of the energy calculations.

Algorithm 2.1: Simplest gradient descent.

Input: some start candidate solution: \mathbf{x}

Result: locally optimal minimum: \mathbf{x}^*

```
begin
  while  $\neg$  endingCrit() do // usually until  $\|\nabla f\| \leq$  threshold of subsequent
     $f$ -values do not change
3   |  $\mathbf{x} \leftarrow \mathbf{x} - \alpha \nabla f(\mathbf{x})$  // step in negative gradient direction with length  $\alpha$ 
    end while
     $\mathbf{x}^* \leftarrow \mathbf{x}$  // found minimum
  return  $\mathbf{x}^*$ 
end
```

previous steps, whose step directions before the current step are not taken into account; the negative gradient direction is pointing in the steepest downhill direction locally, but usually not “globally”, *i.e.*, not to the minimum of the convex region itself. Overall, the convergence rate is therefore quite slow (*i.e.*, linear),^[139] and the method usually descends very slowly when the stationary point is approached.

When Line 3 of Algorithm 2.1 is changed to $\mathbf{x} \leftarrow \mathbf{x} - \alpha[\mathbf{H}_f(\mathbf{x})]^{-1}\nabla f(\mathbf{x})$, *Newton’s method* (for optimization) follows, with the Hessian matrix $\mathbf{H}_f(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ of all partial second derivatives. As a second-order method—that is derived from a Taylor series approximation to second order assuming a twice continuously differentiable function—this usually converges faster, *i.e.* quadratically. If the real function were such a quadratic function, only *one* step would lead to the exact minimum. Tackling more complex non-quadratic functions, at least the implicit second-order approximation becomes better when approaching the minimum in each step. If the Hessian is, however, not positive definite anywhere, the overall step direction could lead to other points.^[139,141,142] A positive definite matrix would have only positive eigenvalues. The type of a stationary point can be defined from the eigenvalues of the Hessian, which are strictly positive for a minimum. Starting far off the minimum, the positive definiteness is not sure and the real function around the minimum might be very badly approximated (*e.g.*, at a non-convex or flat region). In such regions, also a sign inversion due to the inverse Hessian application could impose steps with an *increase* of the function value, which could result in the convergence to the nearest (non-minimal) stationary point, or even to no convergence at all. Moreover, when some eigenvalue approaches zero, the step-length can be out of bounds if no improved step-length schemes are used, which is indicated by the α . There are different methods for setting this step length and other sub-variants of this Newton method, including exact or non-exact line searches, which also can control the convergence behavior.^[139]

What makes things worse is the inversion of the Hessian in each step. It requires the partial second derivatives to be computationally feasible, *i.e.*, they should not be numerically evaluated by finite differences in practice. Additionally, the inversion (or diagonalization) might become computationally expensive. Thus, for commonly convex functions with a proper minimum, the state of the art schemes for a local optimization are the *quasi-Newton* methods: An approximate Hessian (or directly its inverse) is introduced that is updated

by accumulating and merging gradient information from the steps beforehand. Hence, quasi-Newton methods usually start from a (scaled) gradient descent step and “learn” the curvature information encoded in the Hessian in the subsequent steps. Again, there are multiple variants of update schemes available. The most prominent one is **Broyden-Fletcher-Goldfarb-Shanno (BFGS)** or its extension **limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS)** (cf. Ref. [139] for both algorithms) without using a full $\mathbb{R}^{n \times n}$ matrix for the Hessian approximation, but an implicit representation of it.

Still, there are two additional deficiencies: First, the gradients of the function f might not be known at all. For smooth functions, one could of course numerically differentiate the function and use the aforementioned methods. However, specialized derivative-free^[143] optimization algorithms are usually more efficient. In this regard, BOBYQA^[144] and NEWUOA^[145] were used for the present work for bound and unbound problems, respectively. Instead of using a line search procedure, these methods follow a *trust region* approach. In this case, a quadratic model is imposed for a *subset* of the objective function around the current \mathbf{x} by interpolation between calculated values of $f(\cdot)$.^[143] From this, the direction in the higher dimensional space can be inferred and the minimum can be found in the subsequent quadratic sub-problems. A step can be taken and the trust region can be expanded if the model is fine with respect to the real objective function value and contracted if the model is poor, until convergence (for a less-simplified depiction, Refs. [143–145] can be consulted). Secondly, and more important with respect to all the chemical problems of this Thesis, all these aforementioned algorithms only converge to one, usually the nearest, *local* minimum. Thus, *global* optimization approaches are needed. Moreover, methods for arbitrary non-differentiable functions are also required as the property landscape of the inverse design usually brings in this complexity. Such depictions are given in Section 2.1.2. Next, non-linear minimizations using equality and inequality constraints in the form needed for this work are briefly discussed.

2.1.1.1 Penalty Function Method

Constrained optimizations with many different equality and inequality constraints are hard to solve in practice. A well-known methodology is the definition of a higher-order function for the *Lagrange multipliers* method, which can be generalized to work for inequality constraints, by using, e.g., the Karush–Kuhn–Tucker approach.^[139] Usually by imposing all the constraints, not even a minimum is the solution but a critical point in general that can also be a saddle point of arbitrary order, which would lead to the need of specialized solvers. Note that during the global optimization of **GOCATs**, the objective function can be rugged, multi-modal, deceptive, etc., without having analytical derivatives (which would also be impossible at the aforementioned non-differentiable points). These are intricacies of the objective function that are also discussed later in Section 2.1.2. Therefore, we use mainly the *penalty function method* for these cases during global and derivative-free local optimization.

Consider a penalty function that is a function of the objective function itself and the constraints, which are written here together as $c(\mathbf{x})$ (compare with the general definition

of Eqs. (2.1) to (2.3) on p. 13)

$$P(\mathbf{x}) = \bar{P}(f(\mathbf{x}), \mathbf{c}(\mathbf{x})). \quad (2.6)$$

To introduce a measure of how severely the constraints are violated, the constraint violation functions are defined (remember the sets of indices of equality constraints, E , and inequality, I)

$$g_i^{(-)}(\mathbf{x}) = \max\{g_i(\mathbf{x}), 0\} = \begin{cases} 0, & \text{if } g_i(\mathbf{x}) \leq 0 \\ g_i(\mathbf{x}), & \text{if } g_i(\mathbf{x}) > 0 \end{cases}, i \in I, \quad (2.7)$$

$$h_j^{(-)}(\mathbf{x}) = h_j(\mathbf{x}), j \in E. \quad (2.8)$$

and compose them to the vector-valued function

$$\mathbf{c}^{(-)}(\mathbf{x}) = \left(g_i^{(-)}(\mathbf{x}), \dots, g_{m_1}^{(-)}(\mathbf{x}), h_1^{(-)}(\mathbf{x}), \dots, h_{m_2}^{(-)}(\mathbf{x}) \right)^T. \quad (2.9)$$

Now $\mathbf{c}^{(-)}(\mathbf{x}) = \mathbf{0}$ iff $\mathbf{x} \in X$, i.e., the point is in the feasible set. Put differently, the constraint violation functions are nonzero if the corresponding constraint is not fulfilled and zero when the constraint is true. Now, the penalty function can be recast into

$$P(\mathbf{x}) = f(\mathbf{x}) + k(\mathbf{c}^{(-)}(\mathbf{x})), \quad (2.10)$$

with a penalty term $k(\mathbf{c}^{(-)}(\mathbf{x}))$ defined on $\mathbb{R}^{m_1+m_2}$ with m_1 inequality and m_2 equality constraints; the function $k(\cdot)$ that wraps the separate constraints must satisfy

$$k(\mathbf{0}) = 0 \quad \text{and} \quad \lim_{\|\mathbf{c}\| \rightarrow +\infty} k(\mathbf{c}) = +\infty. \quad (2.11)$$

One of the oldest and simplest penalty terms was used for this $k(\cdot)$, the quadratic penalty (also called Courant penalty, or L_2 penalty)

$$P(\mathbf{x}) = f(\mathbf{x}) + \sigma \left\| \mathbf{c}^{(-)}(\mathbf{x}) \right\|_2^2, \quad (2.12)$$

with a penalty coefficient σ . The Euclidean norm or more generally the L_2 norm is denoted by $\|\cdot\|_2$ and this norm is always assumed in this Thesis if no explicit index is given. Remember that $\left\| \mathbf{c}^{(-)}(\mathbf{x}) \right\|_2^2$ simply is the scalar product of the composed constrained violation function, alternatively written as

$$P(\mathbf{x}) = f(\mathbf{x}) + \sigma \left(\sum_i^{m_1} \left(g_i^{(-)}(\mathbf{x}) \right)^2 + \sum_j^{m_2} \left(h_j^{(-)}(\mathbf{x}) \right)^2 \right). \quad (2.13)$$

This problem can be solved as a *series* of *unconstrained* minimization problems, where in each cycle the penalty parameter(s) are increased by, e.g., an order of magnitude.⁵ The solution of the first cycle is used as initial guess for the next round. Multiple such successive iterations will (probably) converge to the solution of the original constrained problem.

⁵ In practice, we start even at $\sigma = 0$ and stop at a large value using multiple rough optimization rounds in order to adhere to the constraints after the tight final optimization round.

In the limit of $\sigma \rightarrow \infty$, all constraints hold since these, otherwise, would add an infinite penalty to the function. If all constraints hold, it follows $P(\mathbf{x}) = f(\mathbf{x})$.

2.1.1.2 Concrete Case of Electrostatic Potential Optimization

In particular, one more specific problem to be briefly discussed is the direct optimization of an **electrostatic potential (ESP)**. This is either done during the so-called Canada search steps, discussed later in Section 3.5, or it is used as a standalone single-objective function for the **GOCAT** optimization. In this case, the *penalty function method* described in the previous Section is used for treating all equality and inequality constraints.^[139] In the case of **ESP** optimization, we want to minimize

$$\min_{\mathbf{x} \in \mathbb{R}^{4n}} P(\mathbf{x}) = \sum_J^N \sum_i^n \underbrace{\left(\frac{q_i}{\|\mathbf{r}_i - \mathbf{R}_J\|_2} - \varphi_J^{\text{ESP}} \right)^2}_{\|\Delta\varphi_J^{\text{ESP}}\|_2^2} + \sigma \left(\sum_k^{m_1} \left(g_k^{(-)}(\mathbf{x}) \right)^2 + \sum_l^{m_2} \left(h_l^{(-)}(\mathbf{x}) \right)^2 \right), \quad (2.14)$$

where the first term is the objective function denoting the difference between a reference **ESP**, φ_J^{ESP} , at each core atom, J , and the calculated one by the current Cartesian positioning, $\{\mathbf{r}_i\}$ with $\mathbf{r}_i \in \mathbb{R}^3$, and charge value, q_i , of the embedding consisting of n charges, leading to $4n$ dimensions in total. Consequently, the domain of the objective function is $\mathbf{x} = \{\mathbf{r}_1, \dots, \mathbf{r}_n, q_1, \dots, q_n\}$. \mathbf{R}_J describes the atom coordinates of the J -th atom. In these applications, mostly three frames were used (**R**, **TS**, **P**), which also allows the translation between levels of theory with a varying number of frames of the **MEP** discretization and their coordinates; discussions regarding the **MEP** and its discretization follow later in Section 2.5.1. Eq. (2.14) optimizes the Coulomb potential at selected Cartesian points, which is given here in atomic units and which will be discussed further in Section 2.3.

The equality constraints are usually set as

$$h_1(\mathbf{x}) = \sum_i^n (q_i) - q_{\text{tot}} = 0, \quad (2.15)$$

$$h_{2, \dots, n+1}(\mathbf{x}) = \|\mathbf{r}_i - \mathbf{R}_{J, \text{Voronoi}}\|_2 - d_J = 0, \quad \text{for each charge } i. \quad (2.16)$$

In Eq. (2.15) the total charge of the **GOCAT** should equal the target one, as, e.g., in the neutral case $q_{\text{tot}} = 0$. For Eq. (2.16), a surface must be defined; in most cases a **vdW** surface was used such that $\{d_J\}$ are the atom-dependent **vdW** surface radii. Strictly speaking, the algorithm is slightly more subtle: First, a best match of one charge and a corresponding core atom must be found. If the charge is outside of the **vdW** surface of each atom, $\|\mathbf{r}_i - \mathbf{R}_J\| \geq d_J, \forall \{i, J\}$, one can then compute the minimum difference, $\min(\{\|\mathbf{r}_i - \mathbf{R}_J\| - d_J\})$, to find a corresponding atom J . If a charge is inside of a **vdW** surface, it must be moved back onto this surface again. Even more subtleties are included in the computations for the cases in which two (or multiple) **vdW** surfaces overlap and/or a charge is in a subset of both. Eventually, each charge i is assigned to *one* corresponding atom J , which is indexed by “Voronoi” above

(to indicate some similarity to a Voronoi tessellation, *i.e.*, a space division into sub-regions with the atoms J as centers).

This general procedure is not limited to **vdW** surfaces and can be extended to work with any other surface, such as a sphere (or a ellipsoid, etc.) by using appropriate distances $\{d_J\}$ to the middle of the sphere (or to the closest points on the ellipsoid).⁶

The inequality constraints usually are:

$$g_{1,\dots,n}(\mathbf{x}) = q_i - q_{\text{up}} \leq 0, \quad \text{for each charge } i, \quad (2.17)$$

$$g_{n+1,\dots,2n}(\mathbf{x}) = q_{\text{low}} - q_i \leq 0, \quad \text{for each charge } i, \quad (2.18)$$

$$g_{2n+1,\dots}(\mathbf{x}) = r_{\text{min}} - \underbrace{\|\mathbf{r}_i - \mathbf{r}_j\|_2}_{r_{ij}} \leq 0 \quad \text{for each pair of charges } \{i, j > i\}. \quad (2.19)$$

q_{up} and q_{low} denote the upper and lower boundaries of charge values, respectively, and r_{min} a minimum distance that is to be conserved. Following the logic of inequality constraining, no penalty is added if $q_i \in [q_{\text{low}}, q_{\text{up}}]$ and all distances r_{ij} are greater than r_{min} . Typical values used are $r_{\text{min}} = 1 \text{ \AA}$ and $[q_{\text{low}}, q_{\text{up}}] = [-1, +1] e$.

Additionally, especially for the **vdW** surface optimization, another penalty term, $k(\cdot)$, is defined that creates a steeper gradient around the feasible set, shown here for a calculated distance as input (*cf.* Eq. (2.16) on the preceding page).

$$\Delta d = \|\mathbf{r}_i - \mathbf{R}_{J, \text{Voronoi}}\|_2 - d_J, \quad \text{for any charge } i \quad (2.20)$$

$$k\left(h^{(-)}(\Delta d)\right) = \kappa_1(\Delta d)^2 + w\kappa_1\left(1 - \exp(-\kappa_2(\Delta d)^2)\right). \quad (2.21)$$

This is a mixture of a quadratic term (first part) and a negative Gaussian (second part). In Fig. 2.1, it is illustrated with typical values at the end of the penalty function series. Analytical gradients of all those terms are readily available, allowing the efficient quasi-Newton solver **LBFGS** to be used.

This **ESP** optimization can be used to translate between levels of theory, which is necessary because the final embedding, *i.e.*, the exposed surface, will usually be different between the levels (see the later application in Section 6.3.2). Moreover, this can generally also include translations on the same level of theory by changing some of the model assumptions, for instance, the number of charges, N_{Ch} . This can be leveraged as a compression technique in order to condense complex **GOCATs** by finding simpler (smaller) ones that are defined by different model settings or boundary constraints.

Note that fitting charges based on **ESP** is a well studied topic in computational chemistry in the context of **potential derived (PD)** atomic charges (*e.g.*, see Refs. [146–148]). However, in contrast to the **GOCAT** design, the problem tackled in the literature is “inverse”: For calculating **PD** charges at *the core atoms*, these are optimized to be the best representation with regard to some calculated molecular **ESP** on an outer surface just beyond the **vdW**

⁶ More precisely, there are two different variants available: First, an arbitrary space as feasible region itself can be used, *i.e.*, by directly only sampling \mathbf{x} with already holding constraints. This is based on numerical tessellation routines for constructing a **vdW** surface (or a **solvent accessible surface**, for instance). Second, generic penalty functions were implemented to add those penalty function values and derivatives for any local optimization.

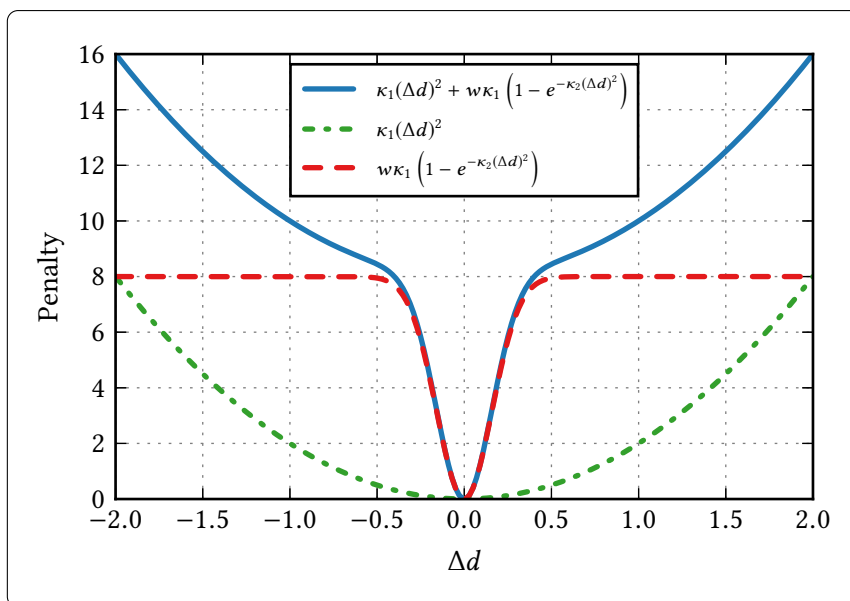


Fig. 2.1: Penalty terms of Eq. (2.21) with $\kappa_1 = 2$, $\kappa_2 = 20$ and the “mixing” weight $w = 4$. Both separate terms—quadratic and negative Gaussian—are also shown. Note that these penalties can generally work on any generic property such that Δd can sign any difference to an intended restraint value. Hence, no physical units for the resulting penalty function value or its parameters are shown.

surface or on a **solvent accessible surface (SAS)** surface. Here, only $\{q_i\}$ are fitted with pre-determined coordinates, namely the atomic positions. Thus, such a problem is of a *linear* least-squares type. However, it commonly entails constraints and/or restraints conserving the total molecular charge, the charge within a molecular part, the overall symmetry or other properties, such as dipole moments. Nonetheless, it is usually an *over-determined* linear system of equations which, in fact, can be highly correlated^[149–151] and thus lead to statistically *underdetermined* charges and hence a rank deficiency of the respective least-squared matrices, especially for large systems.⁷ The underdetermined character of the charges due to the high correlation is intuitively understandable,^[141,152] as the **ESP** outside of the molecule is primarily dependent on the charges near the boundary to the common exposed surface and not as much dependent on the interior (compare with the linear superposition principle in Section 2.3, which also implies in this case that charges can be *shielded*). In this regard, even more restraints are often added in order to simplify the, in fact, underdetermined or ill-posed problem. This is done by incorporating, *e.g.*, (what would nowadays in the **machine learning** community be called) L_2 (or Ridge) regularization terms, besides others, which results in the **restrained electrostatic potential (RESP)** model.^[153,154]

In the context of partial charge optimization of a **GOCAT**, this topic will be discussed again in Section 6.3.2. Next, global optimization is briefly introduced.

⁷ Besides, other problems can emerge, *e.g.*, electron density errors (*i.e.*, regarding the level of theory), basis set dependent partial charges, conformational dependence, sampling density dependence, etc.

2.1.2 Global Optimization

For the purpose of developing some intuition about hard characteristics of functions that frequently appear in practice and demand a global optimization, a qualitative sketch of such complexities is given in Fig. 2.2 on the following page.⁸ These are different quality measure surfaces (or fitness/objective function surfaces), which are plotted each in *just* two dimensions. The target in every case is to find the global minimum robustly, *i.e.*, the lowest value of $f(\mathbf{x})$, irrespective of the starting point: In Fig. 2.2(a), the overall convex case is shown, where even a simple (negative) gradient following algorithm would already find the one and only local *and* global minimum, using the gradient descent in Algorithm 2.1, for example. In Figs. 2.2(b) and 2.2(c), the functions possess multiple minima—the fitness surface is not unimodal anymore—here with less and with more variation, respectively. Consequently, there might be an obviously worse minimum with a high $f(x)$ (Fig. 2.2(b)) or multiple minima that are competitive (Fig. 2.2(c)), separated by large barriers in-between. Increasing the latter feature even more leads to the picture of Fig. 2.2(d), where a highly varying, multi-modal, rugged surface is shown. In this case, local information (such as gradients) will not really help and the optimization could directly get stuck in the nearest optimum. Such *misleading* local information could also be more emphasized in the clearly *deceptive* landscape of Fig. 2.2(e), where, over a longer width, such information is monotonously conducting the search but does not lead to the best minimum. In contrast, a region without any information at all is sketched in Fig. 2.2(f). Often, there are also cases with minor variations and local optima but with a very narrow “needle-like” minimum (Fig. 2.2(g)), or, in Fig. 2.2(h), there are numerous peak-like minima without any local information whatsoever in the proximity of the “needle” (or misleading information at the right-hand side).

Note that these are simplified sketches in *only* two dimensions of an N -dimensional search space. In principle, if each dimension had (just) 10 possible (discrete) values, in a space of N dimensions this would directly lead to the exponential scaling of $O(10^N)$. Naturally, in such continuous search spaces, there are more than just 10 input values for each dimension. This is the intimidating “curse of dimensionality” (a term coined by BELLMAN)^[157] that re-appears in many types of problems.^[29,158,159] However, this does not *necessarily* lead to an exponential scaling of the local minima of the objective function. As a matter of principle, there might be significantly fewer *promising* regions in the search space than this combinatorial number.⁹ If there were just one overall global convex region, the problem would be almost trivially solvable. The same would be true if the problem in N dimensions could be separated into N 1D sub-problems—without any correlations or non-linearities between the dimensions. However, one generally cannot assume this to be true, and for the universal case, global optimization algorithms must deal with such hard optimization problems—or put differently, they, *ipso facto*, are only then needed.

⁸ When not explicitly cited otherwise, this Section is based on the general textbooks and references on this topic.^[29,30,155,156]

⁹ In the well-investigated regime of global structure optimization (of clusters), it was indeed found that not only the abstract combinatorial space scales exponentially but also the physically meaningful space, *i.e.*, the number of local minima and thus the number of structurally *stable* solutions.^[160–165]

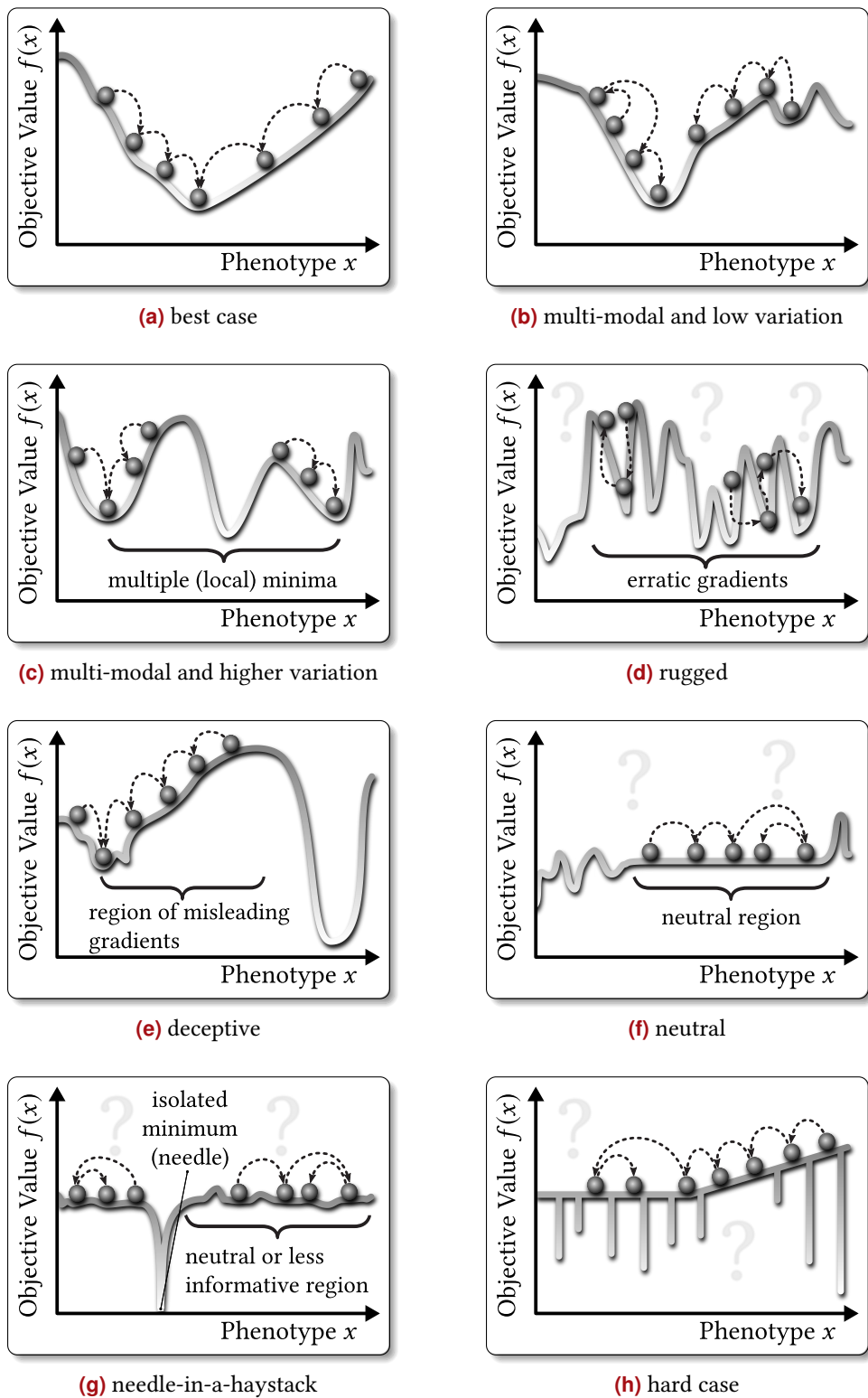


Fig. 2.2: Picturesque illustration of several characteristics of the objective functions, also called fitness surface/landscape (below). Some discrete phenotypes (these terms are explained later, *cf.* Section 2.1.2.2 on p. 27), x , are shown and some possible movements by the arrows. See the main text for further explanations. The pictures are adapted from Refs. [29, 155].

One can divide (global) optimization algorithms into two classes: Exact methods that must (somehow) prove to have found the global minimum and heuristic methods that *lack* such a guarantee.^[166] Often, problems that are faced in practice are either proven to be of the *NP-hard* type, *i.e.*, then it is generally *assumed* that they cannot be solved in polynomial time, and this suggests the exponential scaling of the time needed with the problem dimension for an exact solution.^[29] Similarly, problems are assumed to be of such a type when treating very high-dimensional ones with maybe no simple objective function or prior information that could be exploited. Indeed, in the category of *deterministic* exact approaches, there is also a plethora of techniques available (such as, *e.g.*, *branch and bound* methods) for eliminating many solutions in each iteration and thus shrinking the search space effectively. At the end, each point in search space must be visited somehow or must be proven to be worse than the found minimum.^[166] In other words, the main challenge of each global optimization technique finally is to provide a way of somehow *not* having to calculate each possible point in search space. However, the following discussions about global optimization will focus only on the latter category of optimization algorithms that comprises heuristics that can generally treat any problem without having prior knowledge of it. Accordingly, black-box optimizations are possible which can find a good but maybe not best solution quickly, robustly and often non-deterministically.¹⁰ In contrast, a deterministic algorithm will always return with the same output when the same input is given, whereas in the non-deterministic algorithms, often at least one internal state transition is made probabilistically and they, hence, can return with a different output despite the same starting conditions (input). For those deterministic global optimization methods, the reader is referred to Refs. [168, 169].

Generally, the corner cases or baseline approaches each *more sophisticated* optimization algorithm in the following must prevail are, on the one hand, a fully random (non-deterministic) heuristic search and, on the other hand, a brute-force (deterministic) exact search. The latter, as motivated above, is simply not affordable in practice as long as no shrinkage algorithms or heuristics are used as such an exhaustive enumeration would merely test each solution without exploiting *any* information gained from the objective function. This *trivially* leads to the exponential scaling because of the high-dimensional search space and the encountered *NP-hard* problems in usual real-case applications.^[29] Of course, both approaches could simply be stopped at any time and the best solution returned that has been found so far. Clearly, more efficient and effective algorithms should be available.

So, any robust optimization scheme that tries to find the global minimum must introduce the following qualitative features with respect to the “landscapes” of the objective functions^[166]

- quickly find local optima,
- quickly leave local optima (again),

¹⁰ Again, one should emphasize that although the hardness of cluster optimization was shown,^[161,162] such problems can be solved *heuristically* and non-deterministically in polynomial time, yet generally without the guarantee of having found an exact solution.^[167]

- not get lost in inessential regions (or get misled by deceptive directions),
- and *not* explore the full available search space, but use any type of (meta-)heuristic procedure to explore and refine *promising* regions exploiting information accumulated during the search.

In this regard, one must use some type of optimization algorithm which varies between the two sides of a dichotomy, *i.e.*, between *exploration* and *exploitation*. The former describes the need to investigate new, diverse (promising) regions in search space, including not being stuck anywhere. Exploration usually induce long-distance leaps trough the search space to get far away from the already acquired information to new pieces of information. In contrast, the latter denotes a principle to re-use the candidate solutions' (intrinsic) information acquired so far, also including the *local* information mentioned above, to enforce the progression of the overall optimization. This exploitation can usually be mapped to small steps in the search space.¹¹ The line between those two principles is not clear-cut, which will become apparent when further algorithmic details are considered below. In their implementation, usage and configuration, such algorithms, however, usually reside somewhere in-between these two poles.¹² Again as a limiting case,^[156] a simple gradient descent or Newton's method (*cf.* Section 2.1.1), which is utilizing even more local (curvature) information, could be dubbed fully locally exploitative. On simple surfaces including derivative information (as in Fig. 2.2(a) on p. 22), an algorithm of this kind is the best choice. A full random search, on the other hand, without any history would be (globally) explorative and resistant to ruggedness as in, *e.g.*, Fig. 2.2(d) on p. 22. The real feat, therefore, is to find a balanced setting for the problem at hand.

2.1.2.1 Metaheuristic Optimization

This opens up the way towards a type of “twilight-zone”, the *metaheuristic* optimization. We follow the definition of this particular term which is originally coined by GLOVER^[170] and quoted here from Refs. [25, 26, *emph. in original*]:

A *metaheuristic* is a high-level problem-independent algorithmic *framework* that provides a set of guidelines or strategies to develop heuristic optimization algorithms. The term is also used to refer to a problem-specific implementation of a heuristic optimization *algorithm* according to the guidelines expressed in such a framework.

¹¹ However, note that typical crossover steps that can completely exploit the information of two candidate solutions can lead to huge jumps in search space ending at a combined point of the starting solutions. Therefore, the extent of information that *is* or *is not* re-used is more important than (translated) step lengths.

¹² Intuitively, the following holds (the terminology will become clear below): On the one hand, more exploitative settings include more crossover, smaller population size, local optimizations in-between, less severe niching, more severe selection functions and less diversity checks. On the other hand, more explorative settings utilize bigger/more frequently mutations, bigger populations, severe niching and diversity checks as well as less severe selection. Though, as many operators themselves live in a “twilight zone”, *e.g.*, many crossover operators can also induce giant jumps landing on new (unseen) regions, the mapping of operators to the two-poles is not clear-cut.

This means that, as already stressed in those same references, the term “metaheuristic” has two (different) meanings: Both for denoting a problem-independent algorithmic framework (as *e.g.*, OGOLEM itself for solving specific chemical global optimization problems, detailed in Chapter 3) as well as for denoting specific algorithmic flavors such as, *e.g.*, hill climbing, GA, particle swarm optimization, differential evolution, ant colony optimization, etc. These algorithms often share many commonalities. Most of these different approaches will, however, not be explicated here (see Refs. [26, 29]).¹³

Metaheuristics are often able to work under weaker assumptions such as having no derivatives or even no known or clear form of an objective *function* at all, *i.e.*, they work under *any* quality assessment procedure. In practice, it would suffice to use, for instance, a semi-automatic experimental feedback^[172] or just any simulation cycle such as a robot movement protocol/simulation in order to reach a target. Furthermore, the candidate solution can be of an arbitrary type or representation, not just real-valued vectors in metric spaces, but integer/boolean vectors, encoding maybe just unordered/categorical sets of objects, graph-structures, a catalog of rules, etc.^[156]¹⁴

For its educative value, we take a short look at *hill climbing*—which is in the present minimization context rather an “inverted” hill—as one of the simplest metaheuristics. It is given in Algorithm 2.2 on the next page. By simply drawing a (pseudo-)random number and doing a small, rather local random change to the candidate solution, x , in Line 3 of Algorithm 2.2, this can, by chance, lead to an improvement, resulting in a better quality, which is evaluated next in Line 4. Certainly, how to make such a small modification to x (of arbitrary representation in different problems), is quite another issue. This illustrates the usual strategy of a *resampling technique* (or a “trial and error” approach) where new solutions are generated based on older results, in this case, even without knowing any gradient or direction. When one solution is found, irrespective of where it came from, its quality can be assessed. The usual assumption would be that local changes to x lead to similar (but better) results, x' , which is also called *causality*. Consequently, using such strict local sampling, the overall optimization would still be subsumed as a local one. If this is not appropriate in cases when dealing with multimodal or even highly rugged surfaces, several types of improvements can be made to this simple setting, as for instance: multiple changes to x could be made in one chunk, emulating a gradient descent *via* random steps, and the best x' could be returned. Bigger changes (maybe even re-starts in between) could be incorporated, requirements on the quality of x' could be loosened to allow for (intermediate) worse solutions than x before (Line 4). Finally, maybe even a *whole set* of $\{x\}$ could be treated and processed at the same time—which will be called “population”, see Section 2.1.2.2—and one could strictly remember the best current solution that has been found without replacing this in any other intermediate step.^[156]

Reaching some chemically motivated ingredients, one could not just include binary

¹³The Greek prefix “meta” itself is just partially meaningful and in some cases it might also be thought of a higher-abstraction layer for heuristics, as a heuristic about heuristics; alternatives from Ref. [171] could be used instead such as a *weak stochastic search* or *black-box optimization*.

¹⁴Consequently, metaheuristics can always be used as “last-ditch” methods if no other known technique works.^[156]

Algorithm 2.2: Simplest *hill climbing*.

Input: some candidate solution: x

Result: better candidate solution: x

```
begin
  while  $\neg$  endingCrit( $x$ ) do
    /* maximal iteration counter or quality as threshold (if known) */
3     $x' \leftarrow$  randomChange( $x$ ) // some local change in any/some dimensions
4    if qualityAssessment( $x'$ )  $\leq$  qualityAssessment( $x$ ) then
      /* remember a better solution, otherwise just continue */
       $x \leftarrow x'$ 
    end if
  end while
  return  $x$ 
end
```

if-checks in Line 4, but maybe a chemically motivated probability of acceptance. This could be the well-known Metropolis probability^[27] by using energies as quality, and by including another protocol of changing this MC step acceptance probability gradually throughout the optimization, one reaches the simulated annealing.^[28] Instead, by including *hybridization* protocols into the quality assessment, e.g., by local optimizations in problems where gradients are available such as in **cluster structure optimization (CSO)**, this results in “MC steps with minimization”,^[173] which is also called basin-hopping in this context.^[174,175]

With a few other improvements—but still sticking to the same simple framework picture—finally the **GA** as a metaheuristic is reached that will be discussed below. In the following, the only direct information available to the global optimization algorithm will be the evaluation of $f(x)$ itself—often neither gradients nor any other auxiliary knowledge will be available. This allows for arbitrary rugged (discontinuous) objective functions to be minimized by applying probabilistic transition rules between the solutions. The indirect (secondary) information available can be traced back to the so-called *population*, i.e., the set of multiple concurrent candidate solutions that share their information, which conserves information gained and accumulates some “history”. In this regard, the following descriptions will mainly be limited to one global optimization algorithm family, which is the non-deterministic meta-heuristic pool-based global optimization.

Usually, the terminology in the broad context of **EC**—itself also a generic concept for multiple different stochastic/non-deterministic population-based metaheuristics—(mis-)use terms “inspired” by biology, genetics and evolution.^[25] Hence, **EC** algorithms are commonly motivated and introduced by its similarity to the process of biological evolution. Also the **GA** belongs to this class. Yet, any quality of a search algorithm in computer science should by no means be inferred or semantically transferred from any loose inspiration source. Therefore, as a little disclaimer, though, we will *not* follow any distorted view on **GA** comparing with statements of “how nature optimizes” and will not obfuscate things by motivating that the former “mimics” the latter. Instead, we will simply stick to the mathematical and meta-heuristical foundations.^[176] Notwithstanding, in the literature as well as in this Thesis

(and implementations), some of these terms often slip through. They are introduced in Table 2.1 on the following page and can simply be reduced to the more general neutral terminological counterparts.¹⁵

2.1.2.2 Genetic Algorithm and Terminology

A schematic GA pseudo-code implementation is given in Algorithm 2.3 on p. 29, for which the concepts introduced by Table 2.1 are essential. In the usual work case, a population, P , is randomly generated *via* a full random sampling, *i.e.*, a nullary initialization (Line 2 of Algorithm 2.3). Then, mostly up to a preset maximum iteration number, the shown operators are executed during the main loop. First, some candidate solutions must be selected from the pool that are to be worked on in the iteration (Line 4). This happens in relation to the current fitness of the individuals. The most traditional variant is the *fitness proportionate selection* or *roulette wheel selection*: Each individual is drawn based on its current contribution to the totally summed fitness in the population, with a probability $p(I) = f(I)/\sum_{I \in P} f(I)$.¹⁶ These candidate solutions are changed then by the binary (crossover) and unary (mutation) operators (Lines 5–6), where the latter is not always applied, but to a smaller probability. Simplistic schemes of such operators are shown in Fig. 2.3. The quality of the generated candidate solutions is evaluated, *i.e.*, the objective function (or fitness function) is applied (Line 7). Within the `postSelection(·)` procedure (Line 8), the created children, I'_i, I'_j , might get added to the pool if they fulfill certain requirements such as having a good fitness,

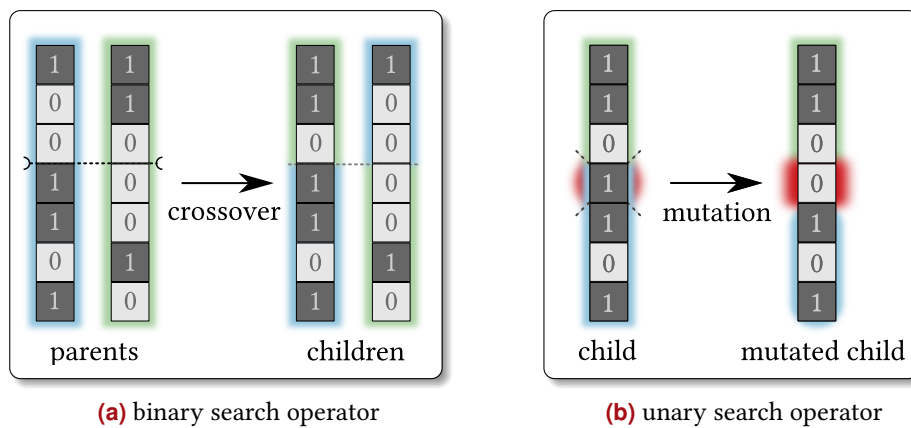


Fig. 2.3: The most traditional GA operators are illustrated. The shown crossover operator divides the genotype into two parts and recombines them, whereas the mutation operator changes a specific information. These simple operators work on binarily encoded genotypes, which leads to the question of how to encode the problem at hand to result in this representation. For all GA applications in this Thesis, the operators will be much more complex and work in a less “distorted” representation. Some of these operators will be discussed in Chapter 3.

¹⁵In many frameworks nowadays, boundaries between elderly distinct metaheuristics are blurred so that a general, neutral terminology would also help to compare the actual algorithmic ingredients and steps, as also discussed in Ref. [176].

¹⁶For translating lower fitness values to be better in minimization problems, some re-weighting of a smaller fitness value to map to a bigger probability for the selection must take place, accordingly. Normalization usually is also very important for not introducing artificial bias into the selection algorithm.^[29,156] Alternatively, non-parametric selection operators are used and discussed in Chapter 3.

Table 2.1: Overview of some terms used in this Thesis (based roughly on Refs. [29, 156, 176]). Note that there are even more terms and sometimes other definitions or confusion about the actual semantics, which will not be recapitulated here.

Nature-inspired Term	Neutral Term	Remarks
genotype	point in search space	denotes what the algorithm operates on (internal/encoded data structure) ^a
phenotype	point in solution space, (candidate) solution	denotes a (decoded) solution in the feasible set and is the input of the objective function
individual	individual ^b	combining genotype and phenotype (mostly used synonymously with phenotype or candidate solution)
selection (sometimes mating)	selection	picking individuals based on their quality
fitness	objective function value	quality after the assessment procedure of the individual/phenotype as input ^c
generation	iteration	our GA itself is generation-free (<i>cf.</i> Chapter 3); usually, one has to define which chain of steps or chunks of operations one iteration comprises
creation of new genotypes	nullary search operation	taking <i>no</i> individual as input in order to generate a new one: <i>e.g.</i> , random initialization
mutation	unary search operation	taking <i>one</i> individual as input and creating a new one based on this template: <i>e.g.</i> , a <i>monte carlo</i> step of some genotype-coordinates in vicinity of the starting point
crossover, recombination	binary search operator	taking <i>two</i> individuals as input and creating (usually) two new ones based on both: <i>e.g.</i> , re-using mixed/merged genotype-coordinates of both parents ^d
child, parent	individual	particular terms for individuals; parents constitute the input for the resulting, operated-upon children

^a Genome, furthermore, often terms the whole search space. Note also that especially the traditional first GAs^[30] worked with binary strings of numbers (termed chromosome then), which was a vital GA characteristic back then. In turn, a function is usually needed that maps those genotypes to phenotypes. In practice, though, especially in the chemical contexts relevant for this Thesis tackling *continuous* problems, this mapping is simply an identity relation, *i.e.*, phenotype and genotype should mean the same thing. In fact, based on the chemical problem, binary representations can be very meaningful representing the quality “on” of “off” for, *e.g.*, the presence or absence in discrete molecular design of functional groups^[177,178] or occupations of sites in alloys.^[179] In general, again following some historical traces, very important improvements were observed when explicit operators with “more phenotype” character were introduced in CSO.^[167,180–182] Therefore, often when more problem-aware operations are used, the term “phenotype” operator was reserved for this.

^b No real alternative is given; in the following, the terms “solution”, or for the explicit problem, *GOCAT*, cluster structure and parameter vector suffice.

^c Generally, there exists a difference between fitness and objective function. The former might be a result of an additional transformation, including *relational* information with regard to the whole population, other heuristics, mapping procedures, etc.^[29] In this Thesis, however, quality, fitness and objective function value simply mean the same thing.

^d Usually, this is supposed to be the most important operation in the GA, mixing meaningful “traits” to reach (hopefully) even better solutions; for the original theories on this such as the schema-theorem, building-block hypothesis^[183] or extended forma analysis, see Ref. [29]. In any case, this assumes what we could call *implicit separability* within the genotype while certain sub-patterns are correlated jointly to the fitness (linkage) in order to reach a better corresponding phenotype. Put differently, if there were no separability of some sub-coordinates, which is *a priori* unknown but nevertheless present, no crossover would lead to systematic improvements of the fitness, only by chance.

Algorithm 2.3: Simple GA.

Data: GA parameters: which (types/chunks of) operators, total iterations, population size, ...

Result: optimized population P^* of individuals $\{I_i^*\}$

```
begin
  /* initialize the population,  $P = \{I_i\}$ , often randomly and/or using older
  solutions (seed) */
2   $P \leftarrow \text{initialization}()$ 
  while  $\neg \text{endingCrit}()$  do // Main GA loop
    /* select 2 individuals, with some prevalence of choosing better ones */
4     $(I_i, I_j) \leftarrow \text{selection}(P)$ 
    /* exchange traits, resulting in 2 new individuals */
5     $(I'_i.g, I'_j.g) \leftarrow \text{recombination}(I_i.g, I_j.g)$ 
    /* by chance: random change of some information */
6     $(I''_i.g, I''_j.g) \leftarrow \text{mutation}(I'_i.g), \text{mutation}(I'_j.g)$ 
    /* quality assessment procedure */
7     $(I''_i.f, I''_j.f) \leftarrow \text{fitnessFunction}(I''_i.x), \text{fitnessFunction}(I''_j.x)$ 
    /* procedure of adding new individual(s) to population  $P$  */
8     $P \leftarrow \text{postSelection}(P, I''_i, I''_j)$ 
  end while
  /*  $P$  might contain the globally best individual now */
   $P^* \leftarrow P$ 
  return optimized population  $P^*$ 
end
```

i.e., a lower value for minimization, and usually some diversity demands. Note that this already illustrates a generation-free GA, as each main cycle creates two individuals and tries to add these to the population, P . Otherwise, the most common implementation is to sample multiple such individuals in order to fill up a new population which replaces the old one partially or completely. The explicit notation using $I.g$, $I.x$ and $I.f$ points to the genotype, phenotype and fitness value of the individual, respectively. Compare with Algorithm 2.2 on p. 26 again, where no selection is needed since no population exists and no recombination can take place, but apart from that, the $\text{randomChange}(\cdot)$ plays a very similar role as $\text{mutation}(\cdot)$ here.

Some notes on the traditionally most relevant implementations for each operator shown could be made—which would already discriminate between the four main sub-classes within EAs, besides others: GA, evolution strategies, evolutionary programming and genetic programming. For further subtle differences, however, the already cited Refs. [29, 30, 156] can be consulted and this historical line will not be followed here: Nowadays, the boundaries between these become increasingly blurred.¹⁷ In this Thesis, we often speak of GA, but as

¹⁷For instance, a GA does not only work on bit-string chromosomes anymore. Often (also in this Thesis), more mutation/exploration is used—similar to the strong/only mutation setting of evolution strategies at the beginning. Indeed, we do not evolve program instructions (genetic programming), but we use internally some graph-based structures as well. Therefore, also these boundaries are not fixed.

multiple ingredients are present that go beyond the traditional framing, the term **EA** is also frequently used synonymously.

This concludes the discussion about optimization for now. Further detailed descriptions of the operators implemented in this Thesis will be re-addressed in Chapter 3, including also the main program package used and extended, namely **OGOLEM**.

2.2 Potential Energy Surface

Treating non-dynamical or time-independent systems in this Thesis, a system of electrons and nuclei is described in non-relativistic **QM** by the **time-independent Schrödinger equation (TISE)**, which is given in Eq. (2.22).^[141,152,184–187] Applying a molecular Hamilton operator, \hat{H} , to a the system, represented by its wave function, Ψ , this leads to the eigenvalue equation for the coupled electronic and nuclear problem

$$\hat{H}\Psi = E\Psi . \quad (2.22)$$

Using atomic units throughout,¹⁸ the molecular Hamilton operator is given as

$$\begin{aligned} \hat{H} = & -\frac{1}{2} \sum_I \frac{\nabla_I^2}{M_I} + \sum_I \sum_{J>I} \frac{Z_I Z_J}{\|\mathbf{R}_I - \mathbf{R}_J\|} \\ & - \frac{1}{2} \sum_i \nabla_i^2 - \sum_i \sum_I \frac{Z_I}{\|\mathbf{r}_i - \mathbf{R}_I\|} + \sum_i \sum_{j>i} \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\|} \end{aligned} \quad (2.23)$$

$$= \hat{T}^n + \hat{V}^{nn} + \hat{T}^e + \hat{V}^{ne} + \hat{V}^{ee} . \quad (2.24)$$

Here, $\{I, J\}$ denote nuclei, $\{i, j\}$ the electrons, $\{\mathbf{r}_i, \mathbf{R}_I\}$ the positions of the separate particles and M_I as well as Z_I the mass and charge of the nuclei, respectively. The operators in Eqs. (2.23) and (2.24) correspond to the kinetic part of the nuclei and electrons (\hat{T}^n, \hat{T}^e), the mutual Coulomb attraction between the electrons and nuclei (\hat{V}^{ne}) as well as nuclear–nuclear and electron–electron repulsions ($\hat{V}^{nn}, \hat{V}^{ee}$). After **BO separation** of this electronic and nuclear problem in Eq. (2.22), the molecular Hamiltonian for the *electronic* part only can be formulated as

$$\hat{H}^e \Psi^e = (\hat{T}^e + \hat{V}^{ne} + \hat{V}^{ee} + \hat{V}^{nn}) \Psi^e = E^e \Psi^e . \quad (2.25)$$

Now, the nuclei positions $\{\mathbf{R}_I\}$ are parameters assuming that the nuclei are quasi-static and hence have fixed values for each atomic structure. As a result, the \hat{V}^{nn} contribution from above is usually added after solving the electronic **TISE** as a simple constant. Furthermore, the actual couplings between nuclei and electrons, now shifted to the nuclear problem (not shown here), can usually simply be *neglected* which leads to the **BO**^[188] *approximation*.¹⁹

¹⁸In atomic units, the Bohr radius, a_0 , elementary charge, e , Planck's constant, \hbar , and Coulomb's constant, $(4\pi\epsilon_0)^{-1}$ (with vacuum permittivity ϵ_0), are all set to 1.

¹⁹This is usually a solid approximation for the electronic ground state and stable configurations (minima) that are very distant to regions where multiple electronic states become energetically similar and couple as in *avoided crossings* or *conical intersections*. The neglected non-adiabatic couplings between separate electronic states would be essential here. Similar thoughts also apply to the further electronic wave function

This corresponds to the quasi-static picture in which electrons are assumed to instantaneously adapt to changes of the nuclei positions. By separating the nuclear and electronic problem, the concept of **PES** follows, where the electronic energy, $E_a^e(\mathbf{R})$, which is given here for the a th electronic eigenstate, acts as a potential for the nuclei. In the next step, the nuclei can be considered to “move on” this **PES**, including the nuclear wave functions that describe, e.g., vibrations, rotations, etc. $E_a^e(\mathbf{R})$ can still be considered as a function of the nuclear coordinates \mathbf{R} . Yet, it is being solved point-wise for each (parametric) nuclear configuration. As usual for electronic structure theory, the actual nuclear dynamics problem is not treated in this work, and we focus on the electronic ground state, $a = 0$, of the molecular system only. From this **PES**, information about the chemical processes such as reactions can often already be deduced and can be statistically described by, e.g., **TST** (cf. Chapter 1).

Next, different methods for calculating the electronic states are briefly discussed, including only those that were used in this work, with some indications to descriptions that can be promising in the future. The following discussions are by no means exhaustive but very selective on this topic, in addressing some important framing of the **GOCAT** design and further basis for the current work. These depictions will also include some empirical potentials as surrogate for the electronic energy surface $E_0^e(\mathbf{R})$, namely **force fields (FFs)**. For **QM** in theoretical chemistry, we refer to the general textbooks on this topic^[141,152,184–187] or the very extensive review^[189] which covers (almost) all methods mentioned here; additional more recent research on specific parts is cited separately.

2.2.1 Hartree-Fock Approximation

As the **TISE** can only be solved for the simplest (one-electron) systems, further approximations are needed. Having to treat fermionic systems, a many-particle wave function should be *antisymmetric* with respect to the permutation of the indistinguishable electrons. Thus, for an *ansatz* for Ψ^e , the most common antisymmetrized description obeying the Pauli exclusion principle is a determinant

$$\Psi^{\text{SD}}(\{\mathbf{x}_i\}_{i=1}^{N_e}) = \frac{1}{\sqrt{N_e!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \cdots & \chi_{N_e}(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \cdots & \chi_{N_e}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_{N_e}) & \chi_2(\mathbf{x}_{N_e}) & \cdots & \chi_{N_e}(\mathbf{x}_{N_e}) \end{vmatrix}, \quad (2.26)$$

which is also known as **Slater determinant (SD)**. Here, N_e electrons occupy the N_e spin orbitals $\{\chi_i\}_{i=1}^{N_e}$. As single-particle functions or one-electron wave functions, these are defined as

$$\chi(\mathbf{x}) = \psi(\mathbf{r}) \cdot \begin{cases} \alpha(\omega) \\ \beta(\omega) \end{cases}, \quad (2.27)$$

with spatial orbital part $\psi(\mathbf{r})$ and the spin functions $\{\alpha(\omega), \beta(\omega)\}$.

This **SD** can now be used as a trial wave function for solving the **TISE** (Eq. (2.22)), which

approximations that are discussed below and that are biased towards single states as in the single-determinant *ansatz* (*vide infra*).

leads to the **Hartree–Fock (HF)** equations by solving Eq. (2.22) variationally. This means the energy calculated by using the **SD ansatz** is an upper bound to the true energy. Generally, it will lack *electron correlation*, or equivalently, it is a mean-field approximation incorporating an average electron–electron repulsion only.

The expectation value of the Hamiltonian, $E = \langle \Psi^{\text{SD}} | \hat{H} | \Psi^{\text{SD}} \rangle$, applied onto the trial **SD**, whose orbitals are then minimized with respect to the energy while conserving the equality constraint of having orthonormal orbitals using Lagrange multipliers, results finally in

$$\hat{f}(1) |\psi_a\rangle = \epsilon_a |\psi_a\rangle . \quad (2.28)$$

Here, the Fock operator, \hat{f} , that consists of the nuclear attraction part, \hat{h} , the Coulomb, \hat{j} , as well as the exchange operator, \hat{k} , are defined as²⁰

$$\hat{f}(1) = \hat{h}(1) + \sum_b^{N_e/2} [2\hat{j}_b(1) - \hat{k}_b(1)] = \hat{h}(1) + \hat{v}^{\text{HF}}(1) \quad (2.29)$$

$$\hat{h}(1) = -\frac{1}{2}\nabla_1^2 - \sum_I \frac{Z_I}{\|\mathbf{r}_1 - \mathbf{R}_I\|} \quad (2.30)$$

$$\hat{j}_b(1) |\psi_a(1)\rangle = \langle \psi_b(2) | r_{12}^{-1} | \psi_b(2)\rangle |\psi_a(1)\rangle \quad (2.31)$$

$$\hat{k}_b(1) |\psi_a(1)\rangle = \langle \psi_b(2) | r_{12}^{-1} | \psi_a(2)\rangle |\psi_b(1)\rangle . \quad (2.32)$$

This is already the *canonical* form after a unitary transformation of the orbitals into the eigenbasis of the Fock operator leading to the **molecular orbitals (MOs)**, $\{|\psi_a\rangle\}$, with orbital energies, $\{\epsilon_a\}$, for *restricted closed-shell HF* (where each spatial orbital is populated by one α and one β spin electron, after integration over spin variables, which is not shown here). “(1)” denotes that all these are effective one-electron operators of one electron in the mean field of all the others, describing the electron–nuclear attraction *via* \hat{h} , the local Coulomb repulsion *via* \hat{j} and the non-local exchange between two electrons stemming from the antisymmetry *via* \hat{k} . Although this resembles an eigenvalue problem, \hat{f} depends *via* \hat{j} and \hat{k} on its own eigenstates, $\{|\psi_a\rangle\}$, and can be solved iteratively in, e.g., a **self-consistent field (SCF)** procedure.

The standard numerical treatment of solving Eq. (2.28) is carried out by introducing a known set of spatial basis functions (usually Gaussian functions or Slater-type orbitals that are still common in semi-empirical calculations) to convert it to standard algebraic equations that can be solved by linear algebra techniques. By expanding each **MO**, $|\psi_a\rangle$, into a set of basis functions $\{\phi_\nu\}_\nu^{M_{\text{basis}}}$

$$|\psi_a\rangle = \sum_\nu C_{\nu a} |\phi_\nu\rangle , \quad (2.33)$$

the **HF** equations in Eq. (2.28) become

$$\hat{f}_a \sum_\nu C_{\nu a} |\phi_\nu\rangle = \epsilon_a \sum_\nu C_{\nu a} |\phi_\nu\rangle . \quad (2.34)$$

²⁰The Dirac-notation^[190] is used for abstract state kets and the scalar product between the linear functional in dual vector space $\langle \psi_a |$ applied to the state $|\psi_b\rangle$ as $\langle \psi_a | \psi_b \rangle = \int \psi_a^*(\mathbf{x}_1) \psi_b(\mathbf{x}_1) d\mathbf{x}_1$.^[191]

Application of a corresponding dual $\langle \phi_\mu |$ leads to

$$\sum_v C_{va} \langle \phi_\mu | \hat{f} | \phi_v \rangle = \epsilon_a \sum_v C_{va} \langle \phi_\mu | \phi_v \rangle \quad (2.35)$$

or, using matrix notation

$$\mathbf{FC} = \mathbf{SC}\epsilon, \quad (2.36)$$

with a diagonal matrix of **MO** energies, ϵ , the Fock matrix, \mathbf{F} , and the matrix of expansion coefficients, \mathbf{C} , in atomic orbital basis functions, which are usually centered at the atomic nuclei and also hence named **linear combination of atomic orbitals (LCAO)**. The used **LCAO** are not orthogonal, leading to the overlap

$$S_{ij} = \langle \phi_\mu | \phi_\nu \rangle \neq \delta_{\mu\nu}. \quad (2.37)$$

Equation (2.36) describes the Roothaan–Hall equations that resemble a generalized eigenvalue problem and can be solved iteratively until an **SCF** solution is reached.

Explicitly, the Fock matrix elements are given as

$$F_{\mu\nu} = \langle \phi_\mu | \hat{f} | \phi_\nu \rangle = \langle \phi_\mu | \hat{h} | \phi_\nu \rangle + \sum_{\lambda\sigma}^{M_{\text{basis}}} D_{\lambda\sigma} \left[\langle \phi_\mu \phi_\nu | \phi_\lambda \phi_\sigma \rangle - \frac{1}{2} \langle \phi_\mu \phi_\sigma | \phi_\lambda \phi_\nu \rangle \right], \quad (2.38)$$

with the *density* matrix elements, $D_{\mu\nu}$, and two-electron repulsion integrals in Mulliken notation, $\langle \phi_\mu \phi_\nu | \phi_\lambda \phi_\sigma \rangle$:

$$D_{\mu\nu} = 2 \sum_{k=1}^{N_e/2} C_{\mu k} C_{\nu k} \quad (2.39)$$

$$\langle \phi_\mu \phi_\nu | \phi_\lambda \phi_\sigma \rangle = \int \phi_\mu^*(\mathbf{r}_1) \phi_\nu(\mathbf{r}_1) \|\mathbf{r}_1 - \mathbf{r}_2\|^{-1} \phi_\lambda^*(\mathbf{r}_2) \phi_\sigma(\mathbf{r}_2) \mathbf{d}\mathbf{r}_1 \mathbf{d}\mathbf{r}_2. \quad (2.40)$$

These matrix elements are addressed below again when further approximations are introduced in **semi-empirical quantum chemistry (SQC)**.

In the limit of an infinite basis set, an error still present in this mean field approach is the *correlation* energy that can formally be defined by the deficiency $E_{\text{corr}} = E_{\text{exact}} - E_{\text{HF}}$. This leads to all types of post-**HF** methods that incorporate both *dynamical* correlations as well as *static* correlations, although this distinction is not always clear.^[192] With regard to the **HF** error, the former describes the lack of *instantaneous* interactions between the electrons, whereas the latter tackles the fundamentally erroneous description of using *one SD* for systems or molecular geometries where multiple such states are needed, as used in *multireference* approaches. As it will become apparent in this work, the use of any post-**HF** method, e.g., Møller-Plesset perturbation theory or coupled cluster theory, are infeasible for **GOCAT** design, due to the high computational costs and the extremely high number of **single point (SP)** calculations of energies and gradients needed (as discussed later in Section 6.3.2). Thus, these approaches and other ones for electron correlations are not

explicated here (*cf.*, *e.g.*, Ref. [193]). The level of theory that is affordable lies at the **SQC** level or below (**FF**). Hence, we focus on these descriptions in Sections 2.2.3 and 2.2.4.

The mentioned *qualitative* deficiency of missing static correlations by using just *one* reference state, the **SD**, could also be important when full reaction paths on the **PES** are described: In the **TS** region, some multi-reference character is often present and, consequently, an appropriate description would improve the **TS** energies for the **GOCAT** design. Here, multiple electronic states or **PESs** can become (near-)degenerate when the electronic state quality suddenly changes, *e.g.*, along bond dissociation curves. However, this topic lies still outside the scope of this Thesis, during these first steps inventing the general framework first of all. Indeed, there are also multiple versions of similar correlation treatments available on an **SQC** level, which were already heavily used for photodynamics simulations in our work group, for example.^[194–196] In the future, an exchange of the level of theory to a more *balanced* description would be straightforward if affordable. Meanwhile for the present work, it is assumed that such energetic deficiencies around a **TS** region are secondary as nowhere *absolute* energetic barriers are of prime interest. Above all, *trends* of relative energetic influences on the **PES** are mainly investigated, and also the parametrization using high-end or experimental data as training set for the **SQC** method can impose some *implicit* correlation energy.

2.2.2 Density Functional Theory

An alternative to wave-function-based approaches such as the aforementioned **HF** is **DFT**, which has become predominant in a plethora of computational research over the last years.

Here, instead of the wave function of $\{\mathbf{x}_i\}_{i=1}^{N_e}$ electron coordinates, the main entity of interest is the electron density which generally is an (observable) property and always depends on just three spatial coordinates, \mathbf{r} ,

$$\rho(\mathbf{r}) = N_e \int \cdots \int |\Psi_0(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_{N_e})|^2 d\mathbf{r}_2 \dots d\mathbf{r}_{N_e}. \quad (2.41)$$

N_e is the electron number. We are interested in the ground state density here, and so the wave function can simply be integrated over all other $N_e - 1$ electronic coordinates but one (due to indistinguishability of the electrons, the numbering does not matter).

The underlying theoretical framework for the molecular properties was established by HOHENBERG and KOHN:^[197] The first Hohenberg–Kohn existence theorem states a relation between the ground state electron density and the so-called external potential $\rho_0 \mapsto v_{\text{ext}}$. In the perspective of **DFT**, electrons interact with each other as well as with an external potential. The external potential for chemical systems is the attraction by the nuclei, which is defined by the nuclear charges and its positions, as usual. When the true (non-degenerate) ground state density is known, the external potential can be deduced and thus the Hamiltonian and the wave function.^[152] The second variational theorem then states that this electron density also obeys the variational principle, similar to **MO** theory

(used above): $E[\tilde{\rho}_0] \geq E_0$.²¹ This means that any trial density, $\tilde{\rho}_0$, will result in an energy which has the true ground state energy as lower bound. In turn, this can as well be used as minimization prescription to find the true density, similar to the HF descriptions above.

Despite the existence and variational nature of the energy as functional of the density, still no one has found a way to generate the true density that yields exact results. The most problematic parts are the precise description of the kinetic energy and the exchange–correlation parts as functionals. Therefore, usual DFT calculations are based on the formalism of KOHN and SHAM (KS),^[198] who reintroduced orbitals (*i.e.*, the wavefunction) into DFT leading to a better description of the kinetic part, but also to one that thus scales with the number of electrons again. At the end, KS DFT is closely related to HF calculations with identical formulae for the kinetic, electron–nuclear and Coulomb electron–electron interactions. Yet, the theory was developed from the point of view of being *exact*, though, the discrepancy to the truth is then shifted to the most important new part, the exchange–correlation functional, for the treatment of all many-body effects that are not present in an HF treatment. By introducing a fictitious non-interacting system of electrons in KS theory, the exact energy functional now reads

$$E[\rho] = T_s[\rho] + V_{\text{ext}}[\rho] + J[\rho] + \underbrace{(V_{\text{ee}}[\rho] - J[\rho] + T[\rho] - T_s[\rho])}_{\equiv E_{\text{xc}}[\rho]}. \quad (2.42)$$

This distributes the total energy functional, $E[\rho]$, into a kinetic part of non-interacting electrons ($T_s[\rho]$), the external potential due to the nuclei ($V_{\text{ext}}[\rho]$), the Coulomb electron–electron interaction part ($J[\rho]$) and the exchange–correlation part ($E_{\text{xc}}[\rho]$). As given in Eq. (2.42), this emphasizes the difference from the true correlated electron–electron interactions, $V_{\text{ee}}[\rho]$, and the true kinetic functional, $T[\rho]$, of *interacting* electrons, which $E_{\text{xc}}[\rho]$ is supposed to correct.

Minimizing this functional of Eq. (2.42) with respect to the (orthonormal) orbitals yields a pseudo-eigenvalue problem for the KS MO orbitals $|\psi_i\rangle$ (compare with Eq. (2.28))

$$\hat{k}^{\text{KS}}(1) |\psi_i\rangle = \epsilon_i |\psi_i\rangle, \quad (2.43)$$

with the KS effective one-electron operator as

$$\hat{k}^{\text{KS}}(1) = -\frac{1}{2}\nabla^2 - \sum_I \frac{Z_I}{\|\mathbf{r} - \mathbf{R}_I\|} + \int \frac{\rho(\mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|} d\mathbf{r}' + \frac{\partial E_{\text{xc}}[\rho]}{\partial \rho(\mathbf{r})} \quad (2.44)$$

$$= -\frac{1}{2}\nabla^2 + v_{\text{ext}}(\mathbf{r}) + v_{\text{coul}}(\mathbf{r}) + v_{\text{xc}}(\mathbf{r}). \quad (2.45)$$

This can also be expanded into a basis set, delivering a similar generalized matrix-eigenvalue problem as in the HF case above (Eq. (2.36)) that can again be solved in an SCF procedure because the electron density is needed for $|\psi\rangle$ and *vice versa*.

Thus, the main endeavor in KS theory now is to find a proper functional for $E_{\text{xc}}[\rho]$, since the true density functional, except for its existence, is unknown. As a result, a vast

²¹ As usual in this context, *functions* are applied to their input variables, whereas *functionals* are applied to functions. The latter are distinguished by using the parentheses “[.]” for signifying the functional type.

number of functionals were developed over the years,^[199] some of which still stem from a solid physical ground, but others also reintroduce empirical parameters and can thus not be strictly called *ab initio* anymore. Some functionals can be ordered with regard to their accuracy, but this is not always the case and, in principle, it lacks such a systematic improvement as the *ab initio* post-HF method tree.^[200] For instance, there are functionals only based on the local density in the **local density approximation (LDA)** variants. Moreover, gradients $\nabla\rho$ together with ρ are used in the **generalized gradient approximation (GGA)**. There are hybrids with some HF exact exchange and more categories (meta-GGA, double-hybrids, etc.). The PBE0 functional^[201] also belongs to this latter category of hybrids and is used in later Sections (Chapter 6). The functional should always be chosen according to the specific problem and should be evaluated on a case by case basis. Then, DFT is usually the most computational efficient method available nowadays, explaining its widespread use. It is computationally of similar expense as HF but includes some degree of dynamic correlation, while still being a single-reference determinantal *ansatz*.

2.2.3 Semi-Empirical Approximation

Global optimization needs a very large number of objective function evaluations, each of which will come with many energy and/or gradient computations of the chemical systems. Thus, less computationally expensive but more approximative methods are necessary. The most computationally expensive part of HF theory is the calculation of all the two-electron repulsion integrals (Eq. (2.40)), as these are the most abundant and scale as $O(M_{\text{basis}}^4)$.

SQC starts from such an *ab initio* HF SCF-MO framework, but introduces drastic further approximations to neglect a huge number of these integral terms; finally, the integral evaluations formally scale as $O(M_{\text{basis}}^2)$.^[202]²² The conceptual ideas behind SQC that also outline the approximations made in the **neglect of differential diatomic overlap (NDDO)**^[203,204] family can be summarized in three points: (1) Only the valence electrons are treated explicitly, while the rest is implicitly taken care of by partially shielded nuclear charges and/or further functions to model the screened repulsions from the implicit core-electrons. (2) Only a minimal basis on each atom for the valence electrons is used, but these are described by Slater-type functions (*s*-function for H, *s*-, and *p*-functions for second and third row elements; though less common, there are also schemes with *d*-orbitals as in MNDO/*d* or PM6/PM7 for higher rows, see below). (3) The basic **zero differential overlap (ZDO)** approximation is applied which sets products between basis functions to zero that describe the same electron but are located on different atoms. By this, all one-electron three-center integrals and all two-electron three- and four-center integrals are neglected and the overlap, $\mathbf{S} = \mathbf{1}$, is a unit matrix. However, the introduced error is supposed to be compensated for by incorporating further analytical functions and parameters that have to be optimized based on experimental and/or higher-level computational data. Note that there exist even more approximated SQC methods such as (the older) **intermediate neglect of differential overlap (INDO)** and **complete neglect of differential overlap (CNDO)** and others that neglect

²² For big molecules, the steps for solving the secular equation and density matrix formation take over, formally scaling as $O(M_{\text{basis}}^3)$.^[202]

even more integrals,^[205] namely two-center two-electron ones. For a historical overview, Ref. [202] can be consulted and these variants will not be further discussed here.

The **SQC** Hamiltonian most often utilized in the present work is the general-purpose **PM7**^[206] method by STEWART. It is based on a *modified NDDO* approximation and a (separately developed) derivative of the original **modified neglect of differential overlap (MNDO)**,^[207,208] to which family also further popular methods belong, besides others: **AM1**,^[209] **PM3**,^[210,211] **MNDO/d**,^[212,213] **PM6**^[214] and **RM1**.^[215] All these methods differ mainly by additional core–core repulsion functions (including also a different number of parameters, either just mono-atomic or diatomic functions), neglected integrals and the optimization of the parameters using different training sets and methodologies. **PM7** is the most recent new parametrization for almost the complete periodic table using extensive experimental and theoretical reference data and also includes dispersion, hydrogen bond corrections and changed electrostatic interactions for modeling solid-state systems, correcting some flaws of its predecessor **PM6** (and the methods before that).^{[206]23}

The exact forms of additional parametric functions are not recapitulated here, but are given in the original literature.^{[206,211,214]24}

Besides the aforementioned line of development of the **NDDO**-family, THIEL and coworkers developed the **OM x** ($x = \{ 1,2,3 \}$) Hamiltonians^[217,218] (and **OMD x** ,^[219] with additional dispersion corrections) as further extensions of the **MNDO** model. These ones were not yet used (for production) in this Thesis, but might be also very promising in the near future. Here, besides also other modifications, orthogonalization corrections by a varying degree were introduced into the one-electron terms of the Fock matrix; **OM2** has the most of these corrections. Usually in the **MNDO**-family without such corrections, Pauli repulsions are emulated by further empirical correction terms in the repulsive core-interactions (since the core electrons are simply missing). In contrast, in the **OM x** methods, the valence-shell and valence-core orthogonality was introduced by effective-core potentials and by further treatment of the resonance integrals. These additional integrals added to the core Hamiltonian part without *explicit* orthogonalization of the basis were derived from a series expansion of a LÖWDIN orthogonalization.^[220] Usually in the **MNDO**-family, the **ROOTHAAN–HALL** Eqs. 2.36 are simply solved as an eigenvalue problem, *i.e.*, $\mathbf{S} = \mathbf{1}$, assuming that the **NDDO** basis functions are already orthogonal.²⁵ **OM x** then is supposed to overcome these defi-

²³ The parametrization cannot be termed “global”: It was a *greedy* multi-sweep optimization of differently weighted parameters corresponding to a block of (more) important elements (H, C, N, O) using the parameters of **PM6** as starting values. In turn, fixing these first optimized parameters, the next blocks of elements could be tackled subsequently until including almost the whole periodic table at the end. This was done in gradient-based optimizations with use of Hessian-information, with line searches on constructed surrogate restrained objective functions (called “perturbation”). For **PM6** (having more parameters than **PM7**), a full-blown global optimization would have to treat about 2,000 parameters in a training set of over 10,000 reference data items, which would indeed be a very ambitious project as already claimed in that reference^[214]—and the present author would agree. In fact, global metaheuristic optimization of special-purpose **SQC** parameters on a smaller scale was done in our work group in Ref. [196].

²⁴ As it was pointed out in Ref. [216], one would need to consult about ten publications for collecting all terms that are incorporated in **PM7**. Thus, Ref. [216] itself can be inquired for the precise functional forms as well as for a critical view on **NDDO** in general.

²⁵ This is done despite the fact that actually overlap integrals are re-introduced into the **MNDO** Fock matrix, but not used for the secular equation;^[220] this is in line with the neglect of two-electron integrals in **NDDO**,

ciencies relevant for a better treatment of conformational properties, hydrogen bonds and TSs. However, parameters are only available for five elements (H, C, N, O, F).^[221]

For the OMx details, the reader is referred to a recent review^[221] about the theoretical framework with all the additional integrals for orthogonalization corrections appearing in the core Hamiltonian (besides the original literature).^[217,218] Additionally, a very thorough benchmark of the aforementioned SQC Hamiltonians was done in Refs. [222, 223] for ground state and excited state properties. Extensive investigations concerning the question which orthogonalization corrections are needed (one electron vs. two-electron, etc.) was also re-evaluated in a big data analysis.^[224] For a critical assessment of the explicit or implicit approximations and assumptions about NDDO-based SQC, see Refs. [216, 225] and Ref. [220].²⁶

A completely separate development line with regard to all the aforementioned SQC methods is the Hamiltonian of GRIMME's GFN-xTB^{[228]27} and its improved GFN2-xTB^[229] method that are based on density functional tight-binding (DFTB),^[230] more specifically the highest (*i.e.*, third)^[231] order of density fluctuations included in the DFTB framework: In DFTB based on the DFT picture, a reference electron density as superposition of neutral atom densities is introduced and the deviation from this reference is expanded into a Taylor series including many integral approximations. Again, the details lie outside the scope of this Thesis.^[220] GFN2-xTB mostly uses global and element-specific parameters only and is parametrized for almost the whole periodic table, but is intentionally biased towards structures, frequencies and noncovalent interactions. As reaction energies are not the main target, these are probably worse represented. In fact, it was recently used in a metadynamics study for conformer, reaction path and compound space exploration^[232]—in the spirit of the highly recognized nanoreactor simulation^[233]—with emphasis on electronically complicated structures as, *e.g.*, open shell species and transition-metal complexes. The structure optimization of large transition-metal complexes was demonstrated to yield accurate results with respect to high-level DFT.^[234] Although the PES descriptions are globally consistent in GFN2-xTB, they are maybe too inaccurate for thermochemical details, including barriers.^[232]

The current author has not yet personally used this level of SQC in the present work. However, as indicated in Section 8.1, there were already some first GOCAT optimizations done on this level, especially for transition-metal complexes. Due to the higher flexibility (of included polarization functions) and maybe a better electrostatic coupling (*cf.* Section 2.4), GFN2-xTB might also be a promising description in the future.

Such SQC calculations are generally three orders of magnitude less computationally

but bigger errors arise from the one-electron integrals. What is more, NDDO itself can be interpreted as an (implicit) *emulation* of a basis transformation of the two-electron integrals.^[216]

²⁶In Ref. [216], it is also pointed out that different kinds of errors, *e.g.*, in simple scaling factors of an integral or in numerical procedures for the search of parameter values, slipped in and are still existent. These flaws were detected not only in the extensive literature on the subject but also in the SQC programs used today, including PM7 as implemented in MOPAC.^[226] However, the parametrization itself was based upon these mistakes and thus the errors are partially compensated. That is, *without* such errors, full new implementations of these MO SQC methods would probably need a new parametrization.

²⁷The full description of the acronym is “*geometry, frequency, noncovalent extended tight binding (GFN-xTB)*”.

expensive than DFT or HF.^[202] They can still describe electron rearrangements, *i.e.*, bond formation and breakage, and they implicitly include some zero point energy (ZPE), correlation and actually even thermal effects (simply due to the optimization based on experimental data). Therefore, this level of theory is the most important one for GOCAT design. However, due to the very empirical nature having many parameters and functional corrections, one must always benchmark or assess the method for the studied problem, and should bear in mind that even systematic errors of fundamentally unknown type during the *global* optimization of chemical systems on this level can sneak in.

2.2.4 Empirical Potential

Lowering the computational expense even more, one reaches the domain of fully empirical potentials, commonly also called FF or MM methods in computational chemistry. As the general goal here is to treat reacting systems, classical non-reactive FFs are useless in this context. However, there are many reactive versions available,^[235] two of which will briefly be described in the following: REAXFF^[236,237] and the quantum mechanically derived force field (QMDF).^[227] where multiple separate FFs of the latter kind are coupled *via* empirical valence bond (EVB).^[238,239] Although REAXFF has not yet been utilized for GOCAT design in this Thesis—however, it could be in the future—an implementation of global optimization using parts of the underlying methodology is explicated in this Thesis later in Chapter 3. Moreover, some parts of the implementations took place during the Master’s Thesis of the present author,^[240] but reached also the present work. This will be readdressed again in Chapter 4.

2.2.4.1 ReaxFF

REAXFF^[236,237] was invented by van DUIN *et al.* as a reactive FF fully based on a bond-order formalism: Just by the current structure of a molecular system, the “bonding pattern” encoded in the bond-order terms is recognized. This includes all distances between atoms of a structure to map the local environment around each atom. In turn, this structure-dependent bond-order is crucial for the final energy terms since each energy contribution except for the non-bonding interactions are made dependent on this. Hence, there are no fixed atom types as in traditional force fields needed that have to define, for instance, specific hybridization states, etc., for the correct bond lengths/angles.^[141] Moreover, as second essential ingredient, a charge-equilibration scheme^[241,242] for a variable charge description is inherently incorporated. The non-bonding terms, including the Coulomb and vdW interactions, are calculated between every single atom—though, they are shielded for small distances—and thus, together with the other bond-order dependent terms, a smooth transition for bond creation and cleavage in the description can be reached overall. First of all invented to describe hydrocarbon systems only,^[236] it was heavily extended over the years to also treat diverse systems (which is best illustrated in Ref. [243, Fig. 2, p. 15011]).

The downside is the higher complexity compared to traditional force fields as now many parameters are needed per atom. Furthermore, some of these lack clear physical meaning,

since they are utilized just as a mathematical ingredient in the functional meshwork and/or are introduced as *ad-hoc* correction to treat specific situations. The parameters are usually highly coupled such that a division into sub-problems of independent chunk-wise optimization is difficult^[244–251] (or needs substantial chemical insight from specialists).^[252] As a result, also a big training set covering many different chemical situations is usually needed where again a one-to-one correspondence between certain reference items and parameters can also be absent (based on the role of the parameters, of course).

To indicate the complexity, the summed energetic expression can be given as^[253,254]

$$E_{\text{REAXFF}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{Coulomb}} + E_{\text{vdW}} + E_{\text{H-bond}} + E_{\text{lp}} + E_{\text{overcoord.}} \\ + E_{\text{undercoord.}} + E_{3\text{-conj.}} + E_{4\text{-conj.}} + E_{\text{triple}} + E_{\text{C}_2} + E_{\text{penalty}} . \quad (2.46)$$

The first five terms in Eq. (2.46) are the usual ones that are also known from traditional **FFs**: Interactions describing the bonds, angles and dihedrals (torsions) and the two non-bonding interactions, namely the electrostatic Coulomb interaction and **vdW** interactions. However, the functional forms are *not* the same as in the traditional **FFs** which typically use harmonic (non-dissociative) terms for bonds and angles, Fourier series for dihedrals, etc.^[141] Instead, *e.g.*, bonding interactions and angles are described by exponentials or Gaussian functions repeatedly scaled by the bond-order factors. What is more, the bonding interaction terms, for instance, do not even have a repulsive part which is mainly delegated to the **vdW** terms in REAXFF. As emphasized above, the bond-order dependency in each of these terms is incorporated here that switches these terms essentially on or off depending on the current structure. This latter point stresses the intricacy of the energy contributions. The next four terms in Eq. (2.46) represent H-bonds, lone pairs and penalties for over- or undercoordination. Finally, the last five contributions describe rather specific situations (what was meant above by *ad-hoc* corrections), such as E_{penalty} to stabilize structures with two double bonds in a row (*e.g.*, allene $\text{H}_2\text{C}=\text{C}=\text{H}_2$), E_{C_2} to correct C_2 -molecule energies, E_{triple} for triple bonds (more specifically, $\text{C}\equiv\text{O}$), and $E_{3\text{-conj.}}$ as well as $E_{4\text{-conj.}}$ for conjugation terms for NO_2 and aromatic species. Yet, the concrete functional expressions are not given here; they are best summarized in Refs. [138, 253].

Naturally, there exist many more reactive **FFs**, also similar bond-order dependent ones, which have been invented over the decades. However, these will again not be summarized here (see the reviews in Refs. [255, 256] and references therein or the general introduction for REAXFF in Ref. [138]). For further prospects on REAXFF and applications so far, the reader is referred to Ref. [243]. One further empirical potential which is *coupled* to be fully reactive is given next.

2.2.4.2 EVB-QMDF

In the **EVB** formalism,^[238,239] two diabatic energy functions describing **PESs**, $\{E_1(\mathbf{R}), E_2(\mathbf{R})\}$, are coupled by a term $C(\mathbf{R})$ that usually is dependent on the structures in coordinate space, \mathbf{R} ; the coupling can also be just a function of energy differences or a constant. In order to create the corresponding adiabatic surfaces again while taking that coupling into account,

mathematically the **EVB** matrix in Eq. (2.47) can be diagonalized, *i.e.*, the secular equation can be solved analytically

$$\begin{pmatrix} E_1(\mathbf{R}) & C(\mathbf{R}) \\ C(\mathbf{R}) & E_2(\mathbf{R}) \end{pmatrix}. \quad (2.47)$$

The resulting eigenvalues represent the coupled adiabatic surfaces, of which the *lower* one represents the new coupled **PES**^[257]

$$E(\mathbf{R}) = \frac{1}{2} \left(E_1(\mathbf{R}) + E_2(\mathbf{R}) \right) \pm \left(\left[\frac{1}{2} (E_1(\mathbf{R}) - E_2(\mathbf{R})) \right]^2 + C^2(\mathbf{R}) \right)^{\frac{1}{2}}. \quad (2.48)$$

Thus, in a region with $C(\mathbf{R}) = 0$, the **PES** is described by the pure (lower energy) surface of the two, $\frac{1}{2}(E_1(\mathbf{R}) + E_2(\mathbf{R})) - \frac{1}{2}|(E_1(\mathbf{R}) - E_2(\mathbf{R}))| = \min(E_1(\mathbf{R}) - E_2(\mathbf{R}))$, whereas in a region around the **TS** the finite coupling generates the smooth transition between the two **PESs**.

There are a multitude of different couplings available: from simple constants that are structure independent up to full Taylor series of $C(\mathbf{R})$ at multiple points with more elaborated coupling functions and additional interpolations. Indeed, current research which exactly tackles this coupling description (besides others) for **EVB-QMDFF** takes place in our work group.²⁸ Hence, the original literature in Refs. [258, 259] gives an overview of other coupling methodologies in the literature as well as of the recently incorporated improvements for **QMDFF** couplings with regard to reaction rate calculations.

QMDFF^[227,260] developed by GRIMME can be understood as special-purpose **FF**. The idea behind **QMDFF** is to reproduce a reference input **PES** as closely as possible near a minimum/equilibrium. Hence, it is specifically fitted to such a reference minimum state, as in the present case of this Thesis separately to the **R** and **P** structures. It uses a fixed set of reference items at the minimum—the structure, frequencies, CM5 charges^[261] and Wiberg–Meyer bond orders^[262,263]—instead of chemically diverse training sets that shall mirror most important chemical situations, as done in other more general-purpose **FF** (*e.g.*, *cf.* REAXFF). It includes also some torsional/conformational degrees of freedom and can reach quite an accurate potential for one chemical system and its potential well around a minimum. Most importantly, it is fully anharmonic and can also describe the dissociation of all bonds which were detected in the reference structure configuration (*via* the bond orders). However, it cannot (yet) describe the forming of *new* bonds in a reaction. The concrete functional terms and parameters are again not given here.^[227]

In order to describe also full chemical reaction(s) around specified **TSs**, the aforementioned **EVB**-coupling is leveraged. The accuracy of the potential fits of the minima was presented already in Ref. [227] and the idea of coupling separate **QMDFFs** with differing difficulty was proposed in Ref. [264].

EVB-QMDFF was used in the present work as lowest-end benchmark **FF** for method developments, including fitness function definitions, operators and other meta-parameters

²⁸ As caveat: **EVB-QMDFF** is implemented in OGOLEM (*cf.* Chapter 3). Yet, further improvements including all of these couplings are developed externally to OGOLEM. Thus, for more complicated reactions for which the simplest coupling is not sufficient anymore these improvements are yet to be implemented.

of the framework. These are partially discussed in Chapter 3, and some benchmarks are given in Section 6.3.1. In that benchmark context, one of the *simplest* coupling methods was still used

$$C(\mathbf{R}) = a \exp\left(-b [E_1(\mathbf{R}) - E_2(\mathbf{R})]^2\right), \quad (2.49)$$

were the two parameters, $\{a, b\}$, were fitted by a Levenberg–Marquardt least squares optimization as detailed in Refs. [258, 265].

2.3 Electrostatics

For later analyses of the optimized GOCATs, in which scalar fields, ESPs, and their vector fields, electric fields (EFs), will be looked at, some short descriptions are in order. Electric charges at *rest* are dealt with in the following. So, the general laws of electrodynamics (Maxwell) do not have to be introduced, but merely some common facts about such a static potential and gradient field that each (point) charge at *rest* produces are considered. Hence, the electric field is conservative meaning that the curl $\nabla \times \mathbf{F} = 0$ is zero as no time-varying magnetic/electric fields with mutual induction are present.

2.3.1 Coulomb's Law

The ESP is the amount of work needed to move a positive unit probe charge, *i.e.*, the work *per* unit charge, from a reference point into the electric field, \mathbf{F}_{EF} , infinitesimally slowly (without acceleration) and thus can be described by the line integral over an arbitrary path from zero field to the current point^[266]

$$\Delta\varphi_{\text{ESP}} = - \int_{\mathbf{r}_A}^{\mathbf{r}_B} \mathbf{F}_{\text{EF}} \, d\mathbf{r} = \varphi_{\text{ESP}}(\mathbf{r}_B) - \varphi_{\text{ESP}}(\mathbf{r}_A). \quad (2.50)$$

When starting at a field free reference (at infinity), $\varphi_{\text{ESP}}(\mathbf{r}_A) = 0$, the ESP at point \mathbf{r}_B can be defined as $\varphi_{\text{ESP}}(\mathbf{r}_B)$. Note that the potential is thus also just defined up to an arbitrary constant.²⁹ Due to the conservative nature of the electric field and hence path independency, $-\nabla\varphi_{\text{ESP}}(\mathbf{r}) = \mathbf{F}_{\text{EF}}(\mathbf{r})$ as (negative) gradient of the potential holds. The electric potential energy follows as $V = q \cdot \varphi_{\text{ESP}}$ of a charge q .³⁰ The electric field arising from one or, as shown here, from a set of discrete partial charges, $\{q_i\}$, is known as the Coulomb potential

$$\varphi_{\text{ESP}}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_i \frac{q_i}{\|\mathbf{r} - \mathbf{r}_i\|}, \quad (2.51)$$

with the Coulomb constant $1/4\pi\epsilon_0$ (with vacuum permittivity ϵ_0), where due to the superposition principle each potential (and field) can simply be added. Henceforth, atomic units are used throughout again, and q_i is therefore given in units of elementary charges. The actual

²⁹In the first investigations of Section 6.2, this led to redundant solutions (*i.e.*, working in a needlessly too large search space) off-shifted by a constant.

³⁰Hence, the dimensions for the ESP are “energy per charge” (Volts), often described in $\text{kcal mol}^{-1}\text{e}^{-1}$ in this Thesis.

Coulomb (squared distance) law is known in scalar form as $F_{\text{force}} = q_1 q_2 / r^2$ and describes the force between two point charges.³¹ Vectorially, again for a set of charges, it follows

$$\mathbf{F}_{\text{EF}}(\mathbf{r}) = \sum_i \frac{q_i(\mathbf{r} - \mathbf{r}_i)}{\|\mathbf{r} - \mathbf{r}_i\|^3} = \sum_i \frac{q_i \hat{\mathbf{r}}'_i}{\|\mathbf{r}'_i\|^2}, \quad (2.52)$$

using the normalized vector from the point charge position $\hat{\mathbf{r}}'_i = (\mathbf{r} - \mathbf{r}_i) \|\mathbf{r} - \mathbf{r}_i\|^{-1}$. In turn, the Coulomb force on a charge q at \mathbf{r} is $\mathbf{F}_{\text{force}}(\mathbf{r}) = \mathbf{F}_{\text{EF}}(\mathbf{r})q$.

2.3.2 Molecule Exposed to a Non-Uniform Electric Field

A molecule in any external ESP is considered, $\varphi_{\text{ESP}}(\mathbf{r})$, where the additional indexes “ESP” and “EF” for its field are dropped during this Section. With this potential, always a corresponding electric field is generated (see above), $\mathbf{F}(\mathbf{r}) = -\nabla\varphi(\mathbf{r})$, whose components will also be abbreviated as $\partial\varphi/\partial r_\alpha = \varphi_\alpha$ using $\alpha, \beta, \dots \in \{x, y, z\}$, and $\partial^2\varphi/\partial r_\alpha\partial r_\beta = \varphi_{\alpha\beta}$ as *field gradient*, etc. Then, the ESP can be expanded in a Taylor series (more concretely McLaurin series) after assuming a suitable origin, $\mathbf{r}_0 = \mathbf{0}$:^[186,267]

$$\varphi(\mathbf{r}) = \varphi(\mathbf{0}) + \sum_\alpha r_\alpha \frac{\partial\varphi(\mathbf{0})}{\partial r_\alpha} + \frac{1}{2} \sum_{\alpha\beta} r_\alpha r_\beta \frac{\partial^2\varphi(\mathbf{0})}{\partial r_\alpha\partial r_\beta} + \frac{1}{3!} \sum_{\alpha\beta\gamma} r_\alpha r_\beta r_\gamma \frac{\partial^3\varphi(\mathbf{0})}{\partial r_\alpha\partial r_\beta\partial r_\gamma} + \dots \quad (2.53)$$

Next, a Coulomb interaction operator is defined

$$\hat{V} = \sum_k q_k \varphi(\mathbf{k}), \quad (2.54)$$

where k is used for each particle carrying a charge q_k (nuclei, electrons), which is described by a Cartesian vector $\mathbf{k} = (k_\alpha, k_\beta, k_\gamma)^T$. With this, the interaction of the molecule with the field can be denoted as

$$\hat{V} = \sum_k \varphi(\mathbf{k}) q_k \quad (2.55)$$

$$= \varphi(\mathbf{0}) \underbrace{\sum_k q_k}_{\hat{M}} + \sum_\alpha \varphi_\alpha(\mathbf{0}) \underbrace{\sum_k k_\alpha q_k}_{\hat{M}_\alpha} + \frac{1}{2} \sum_{\alpha\beta} \varphi_{\alpha\beta}(\mathbf{0}) \underbrace{\sum_k k_\alpha k_\beta q_k}_{\hat{M}_{\alpha\beta}} \quad (2.56)$$

$$+ \frac{1}{3!} \sum_{\alpha\beta\gamma} \varphi_{\alpha\beta\gamma}(\mathbf{0}) \underbrace{\sum_k k_\alpha k_\beta k_\gamma q_k}_{\hat{M}_{\alpha\beta\gamma}} + \dots \quad (2.57)$$

$$= \hat{M}\varphi + \sum_\alpha \hat{M}_\alpha \varphi_\alpha + \frac{1}{2} \sum_{\alpha\beta} \hat{M}_{\alpha\beta} \varphi_{\alpha\beta} + \frac{1}{3!} \sum_{\alpha\beta\gamma} \hat{M}_{\alpha\beta\gamma} \varphi_{\alpha\beta\gamma} + \dots, \quad (2.58)$$

where $\varphi(\mathbf{0}) = \varphi$ was used at the end to simplify notation. Now, the first two expressions can readily be identified as $\hat{M} = \sum_k q_k = \hat{q}$, the zeroth moment, being the monopole (*i.e.*,

³¹ Due to consistency with later chapters, \mathbf{F}_{EF} was used for the electric field, not \mathbf{E} . To distinguish the force an additional index is added.

total charge) and $\hat{M}_\alpha = \sum_k q_k k_\alpha$ as the dipole moment (first moment) component of the α coordinate, $\hat{\mu}_\alpha$. The other moments are a bit more tricky to identify since the elements of the tensor of, e.g., $\hat{M}_{\alpha\beta}$, should have 5 instead of the 9 independent components because of the symmetry of mixed second derivatives $\hat{M}_{\alpha\beta} = \hat{M}_{\beta\alpha}$ and the further relation due to the Laplace equation (that can be obtained from the Maxwell equations), $\nabla^2 \varphi = \Delta \varphi = 0$ in charge-free regions. In this regard, usually the *traceless* Cartesian moments, $\sum_\alpha \hat{M}'_{\alpha\alpha} = 0$, are introduced

$$\hat{M}'_{\alpha\beta} = \sum_k q_k (k_\alpha k_\beta - \frac{1}{3} k_\alpha^2 \delta_{\alpha\beta}) = \frac{2}{3} \hat{\Theta}_{\alpha\beta}. \quad (2.59)$$

The other moments can similarly be transformed.^[267] With this, Eq. (2.58) can be recast as

$$\hat{V} = \hat{q}\varphi + \sum_\alpha \hat{\mu}_\alpha \varphi_\alpha + \frac{1}{3} \sum_{\alpha\beta} \hat{\Theta}_{\alpha\beta} \varphi_{\alpha\beta} + \frac{1}{5 \cdot 3} \sum_{\alpha\beta\gamma} \hat{\Phi}_{\alpha\beta\gamma} \varphi_{\alpha\beta\gamma} + \dots \quad (2.60)$$

The conceptual insight one can gain after these derivations is the statement that the monopole, if present, *i.e.*, $q \neq 0$, will interact with the **ESP**, φ_{ESP} , directly, and the dipole moment, $\boldsymbol{\mu}$, with the electric field, $\mathbf{F} = -\nabla\varphi = -(\varphi_\alpha, \varphi_\beta, \varphi_\gamma)^T$. Then the quadrupole moment is responsible for the interaction with a *change* of the electric field and so on. In general, the lowest non-zero moment will dominate the interaction and will be coordinate-system independent.

Using perturbation theory, one can use the defined interaction Hamiltonian to yield the first-order energy, acting on the ground state

$$E_1 = \langle 0 | \hat{V} | 0 \rangle \quad (2.61)$$

$$= q\varphi + \sum_\alpha \mu_\alpha \varphi_\alpha + \frac{1}{3} \sum_{\alpha\beta} \Theta_{\alpha\beta} \varphi_{\alpha\beta} + \frac{1}{5 \cdot 3} \sum_{\alpha\beta\gamma} \Phi_{\alpha\beta\gamma} \varphi_{\alpha\beta\gamma} + \dots, \quad (2.62)$$

with $\langle 0 | \hat{\mu}_\alpha | 0 \rangle = \mu_\alpha$, $\Theta_{\alpha\beta} = \langle 0 | \hat{\Theta}_{\alpha\beta} | 0 \rangle$, etc. Thus, the energy change due to the inhomogeneous electric field can be seen. Looking at the second-order perturbation, the energy will result in (see Refs. [186, 267] for the derivations that are not given here)

$$E_2 = - \sum_{n \neq 0} \frac{\langle 0 | \hat{V} | n \rangle \langle n | \hat{V} | 0 \rangle}{E_n - E_0} \quad (2.63)$$

$$= -\frac{1}{2} \sum_{\alpha\beta} \alpha_{\alpha\beta} \varphi_\alpha \varphi_\beta - \frac{1}{3} \sum_{\alpha\beta\gamma} A_{\alpha,\beta\gamma} \varphi_\alpha \varphi_{\beta\gamma} - \frac{1}{6} \sum_{\alpha\beta\gamma\delta} C_{\alpha\beta,\gamma\delta} \varphi_{\alpha\beta} \varphi_{\gamma\delta} - \dots \quad (2.64)$$

The first term brings in the *polarizability* tensor, $\alpha_{\alpha\beta}$, which describes the change of the dipole moment due to the field component $-\varphi_\beta$. The second term is a cross-term because of the inhomogeneity of the electric field. That is, $A_{\alpha,\beta\gamma}$ describes the coefficients stemming from integral products of $\hat{\mu}_\alpha$ times $\hat{\Theta}_{\beta\gamma}$ and can be pictured both as the additional dipole induced by the field gradient, *i.e.*, the second **ESP** derivatives, and the additional quadrupole moment induced by the electric field. The last term shown here, $C_{\alpha\beta,\gamma\delta}$, describes the quadrupole induced by the field gradient. From the third perturbation order, terms result

such as

$$E_3 = \frac{1}{6} \sum_{\alpha\beta\gamma} \beta_{\alpha\beta\gamma} \varphi_{\alpha} \varphi_{\beta} \varphi_{\gamma} + \dots, \quad (2.65)$$

where the first (dipole) hyperpolarizability as non-linear moment is given as $\beta_{\alpha\beta\gamma}$.

To sum up, finally it results,^[186] assuming $q = 0$ for the moment (neutrality of the molecule)

$$E = E_0 + E_1 + E_2 + E_3 + \dots, \quad (2.66)$$

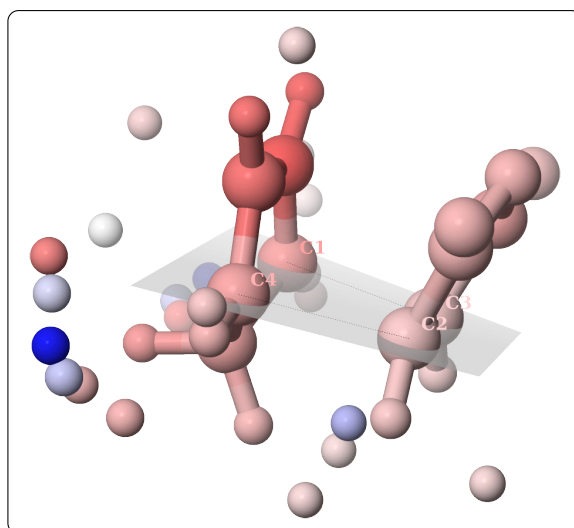
sorting the contributions with respect to the moments gives

$$\begin{aligned} E &= E_0 + E_{\mu} + E_{\Theta} + E_{\mu-\Theta} \dots & (2.67) \\ &= E_0 & \text{(unaffected energy)} \\ &+ \sum_{\alpha} \mu_{\alpha} \varphi_{\alpha} - \frac{1}{2} \sum_{\alpha\beta} \alpha_{\alpha\beta} \varphi_{\alpha} \varphi_{\beta} + \frac{1}{6} \sum_{\alpha\beta\gamma} \beta_{\alpha\beta\gamma} \varphi_{\alpha} \varphi_{\beta} \varphi_{\gamma} - \dots & (E_{\mu}) \\ &+ \frac{1}{3} \sum_{\alpha\beta} \Theta_{\alpha\beta} \varphi_{\alpha} \varphi_{\beta} - \frac{1}{6} \sum_{\alpha\beta\gamma\delta} C_{\alpha\beta,\gamma\delta} \varphi_{\alpha} \varphi_{\beta} \varphi_{\gamma} \varphi_{\delta} + \dots & (E_{\Theta}) \\ &- \frac{1}{3} \sum_{\alpha\beta\gamma} A_{\alpha,\beta\gamma} \varphi_{\alpha} \varphi_{\beta} \varphi_{\gamma} + \dots & (E_{\mu-\Theta}) \end{aligned}$$

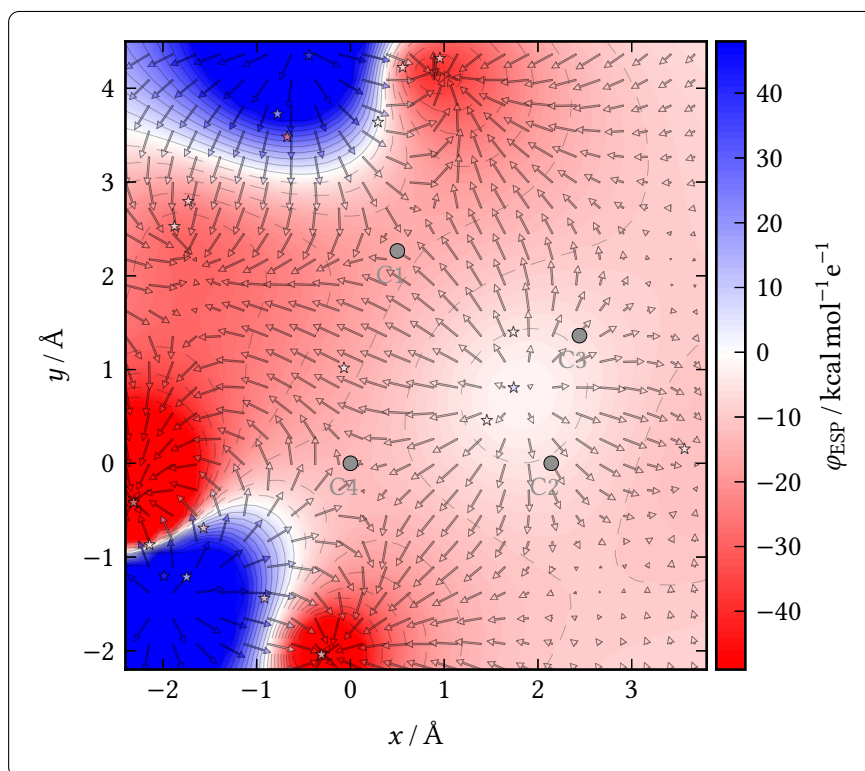
Additional moments, such as the octupole with corresponding polarizabilities and hyperpolarizabilities and their cross-terms, would follow next in the series. In a *uniform* electric field and with a neutral molecule, $q = 0$, the interaction would reduce to the dipole term only. Starting with the permanent dipole term and consequently all its higher-order polarized ones, the energy difference induced by the field would be $\Delta E = E_{\mu}$.

Note that the actual coupling model used within this work is detailed in Section 2.4. There, the energy within the **QM/MM** model that is used and (all) the induction terms that are included by solving the **QM** part variationally will be investigated again. Because a **GOCAT** consists of partial point charges that are mostly sitting nearby the molecules and generating multiple sources and sinks of the **ESP**, they will create a highly inhomogeneous or non-uniform electric field as linear superposition of each charge's **EF**. One arbitrary example is illustrated in Fig. 2.4 on the next page.

A homogeneous field would be the result of some type of (very far away) sitting charges (maybe of the same absolute values but with an inverted sign), as within an (infinite) capacitor from a voltage bias. Every other distribution of charges will always lead to higher order partial derivatives of the **ESP**. Consequently, also an impact proportional to all higher order moments (quadrupole, octupole, ...) is possible. Additionally, all the non-permanent induced moments related to their respective polarizabilities can play a role, and they are all incorporated by the used coupling model already (see below). Thus, for the remainder of this Thesis, Eq. (2.67) will not be used directly to calculate or estimate further energy differences based on *full* perturbation theory. But indeed, these conceptual insights are



(a) TS frame of a Diels–Alder reaction in a GOCAT



(b) ESP and EF in a typical GOCAT

Fig. 2.4: For illustration purposes, the TS frame of a Diels–Alder reaction of maleic anhydride and cyclopentadiene (see Chapter 7 on p. 177) within a GOCAT is shown in Fig. (a) (with a number of $N_{\text{Ch}} = 20$ charges). In the 2D plane of the newly created C-bonds (gray) the ESP and its negative gradient, the electric field, is plotted in Fig. (b) on this 2D plane. There, the “stars” are the partial point charges projected onto the plane that is defined by the 3 C atoms (C4, C2, C1)—these charges are only indicated with small symbols as the main issue is the EF they produce and not the exact positions of the charges. Atoms and charges are colored from red to blue for $q_i \in [-1.0, +1.0] e$ and $\varphi_{\text{ESP}} \in [-47.8, +47.8] \text{ kcal mol}^{-1} e^{-1}$ in both subfigures. The gradients are scaled for visualization purposes by shortening the vectors to a maximal length if they become too large at, e.g., the singularities of the sinks and sources. In contrast to this field here, both the field strength and direction would be the same at each point in a homogeneous or uniform electric field.

needed for both the feature correlations in the later Sections as well as for discussions and comparisons to experimental results. Then, a simple dipole approximation is used. As a recent review on **EF** in structure and reactivity control including catalytic effects, Ref. [268] can be consulted. There, also more involved qualitative changes (based on, *e.g.*, valence orbital theory) are depicted for creating an intuition on how an *oriented* **EF** can influence bonds, structures and reactions. Some of these insights will then be discussed in the respective chapters (*cf.* Section 6.3.3 and Chapter 7).

2.4 Coupling Model

For treating the electrostatic influence of a **GOCAT** on the reaction, we harnessed the already well-known and often used separation into **QM** and **MM** regions (**QM/MM**). Usually, this method is utilized when large systems with many atoms, and more so for **molecular dynamics (MD)** simulations, are to be investigated: The important part of a system with, *e.g.*, a chemical reaction that needs a detailed description, has to be treated quantum-mechanically,³² while the rest of the system is modeled on the lower level of theory, the **MM** methods, since it mainly plays the role of an environment. This can be an explicit solution^[269] or even a full protein surrounding the catalytic pocket.^[270,271] As an extension of this idea, divisions of the overall system into multi-layer systems are imaginable.^[272] The original idea of **QM/MM** can be ascribed to WARSHEL and LEVITT,^[273] who were awarded the Nobel prize in chemistry for their achievements in regard of such multiscale models.

However, for the current work, not the computationally efficient simulation of large systems, but merely the coupling model is of interest here:^[274] The reaction, represented by multiple frames along the **MEP** (see Section 2.5), also including the stationary states, **R**, **TS** and **P**, is treated on a **QM** level of theory, whereas the **MM** is solely made of partial charges, *i.e.*, the abstract **GOCAT** entities themselves. This was treated in a *additive* scheme to embed the **QM** region *electrostatically* (see Refs. [274–276] for other variants and the terminology). In this way, the **QM** region in such a scheme can adapt to the changes in the charge distribution, *i.e.*, the **GOCAT**, and is consequently polarized by it. The Coulomb interaction term and the corresponding energy change is fully treated on a **QM** level. In the following, a short description of this method as well as some remarks are given.

2.4.1 General Quantum Mechanics/Molecular Mechanics

The total Hamiltonian for the composed system is divided into the pure **QM**, the pure **MM** and an interaction part between those

$$\hat{H}_{\text{total}} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}}. \quad (2.68)$$

³²There are reactive **FFs** available (*cf.* Section 2.2.4), but we deliberately exclude a *reactive* **MM/MM** setting.

The coupling Hamiltonian part can usually be described as

$$\begin{aligned}
\hat{H}_{\text{QM/MM}} = & \underbrace{-\sum_i \sum_m \frac{q_m}{R_{im}}}_{\text{attraction of el. \& MM}} + \underbrace{\sum_m \sum_J \frac{q_m Z_J}{R_{mJ}}}_{\text{repulsion of nuclei \& MM}} \\
& + \underbrace{\sum_m \sum_J \epsilon_{mJ} \left(\left(\frac{\sigma_{mJ}}{R_{mJ}} \right)^{12} - \left(\frac{\sigma_{mJ}}{R_{mJ}} \right)^6 \right)}_{\text{vdW interactions of nuclei \& MM}}.
\end{aligned} \tag{2.69}$$

For the **QM** part, i is used for the electrons and J for the nuclei with nuclear charge of Z_J . For of the **MM** part, m counts the **MM** atoms which are, in the present case, the point charges themselves, with charge q_m . R denotes the corresponding Euclidean distances. The last term shows a **vdW** interaction between the outer **MM** atoms and the **QM** nuclei, represented here by a simple **LJ** potential. This potential is defined by the parameters ϵ_{mJ} and σ_{mJ} . However, the partial point charges within a **GOCAT** do not carry any atomic quality (yet). Hence, no **vdW** interactions are used later on, but just the non-bonding electrostatics part is included in the Hamiltonian. Besides that, in the most generic settings, also bonding interactions are needed in Eq. (2.69) (not shown). These must be incorporated if the partition cuts *through* bonds such that there is *not* strictly a non-bonding outer shell and a **QM** inner shell. This would lead to a boundary region of bonded atoms that would need to be treated specially. Such boundary schemes (including linking/connection^[277] atoms, etc.) are not explained here (*cf.* Ref. [274]), since the **GOCATs** so far also follow that simpler category of strict non-bonding interactions.

Then, the calculation of the expectation value of that **QM/MM** Hamiltonian acting on a **SD**, Ψ , yields

$$\begin{aligned}
\langle \Psi | \hat{H}_{\text{tot}} | \Psi \rangle &= \langle \Psi | \hat{H}_{\text{QM}} + \hat{H}_{\text{QM/MM}} | \Psi \rangle + \langle \Psi | \hat{H}_{\text{MM}} | \Psi \rangle \\
&= \left\langle \Psi \left| \hat{H}_e - \sum_i \sum_m \frac{q_m}{R_{im}} \right| \Psi \right\rangle + \sum_I \sum_{J>I} \frac{Z_I Z_J}{R_{IJ}} + \sum_m \sum_J \frac{q_m Z_J}{R_{mJ}} \\
&\quad + \sum_m \sum_J \epsilon_{mJ} \left(\left(\frac{\sigma_{mJ}}{R_{mJ}} \right)^{12} - \left(\frac{\sigma_{mJ}}{R_{mJ}} \right)^6 \right) + E_{\text{MM}}.
\end{aligned} \tag{2.70}$$

This shall illustrate that the electronic part of the Hamiltonian, \hat{H}_e , is appended by the Coulomb attraction of the new **MM** centers, m (first term). Due to the **BO** approximation (see Section 2.2), the nuclear–nuclear repulsion term and the new nuclear–**MM** repulsion term is independent of the electronic degrees of freedom and thus evaluated already—besides the aforementioned **vdW** interactions that are also just dependent on the parametric nuclei coordinates. Note that the only new integrals appearing in the **QM** calculation are, hence, **MM**-atom one-electron attraction integrals, which behave the same as usual electron–nuclear attraction integrals except for the variable partial charge. The charge can also be negative and consequently *repulsion* integrals can result.

2.4.2 Coupling in Purely Electrostatic Globally Optimal Catalysts

For the **GOCAT** model, we neither have any **vdW** interactions to the nuclei nor the energy of the **MM** system is included, denoted as E_{MM} , *i.e.*, the second line in Eq. (2.70) is absent.³³ Neglecting these terms and just including the Coulomb electrostatics of the **GOCAT** leads to following Coulomb interaction energy, given to first order (*vide infra*)

$$E_{\text{Coul}} = - \sum_m q_m \int \frac{\rho(\mathbf{r})}{\|\mathbf{r} - \mathbf{R}_m\|} d\mathbf{r} + \sum_m \sum_J \frac{q_m Z_J}{\|\mathbf{R}_J - \mathbf{R}_m\|}, \quad (2.71)$$

with the (3D) spatial electron density that was already defined in Eq. (2.41) on p. 34. The added essential Coulomb interaction between the **GOCAT** (**MM**) and **QM** system can therefore also be described as the energy of point charges, q_m , interacting with the (fuzzy) **QM**-based **ESP**. This can be pictured as interaction between the charge at position \mathbf{r}_m and the net (partially compensated) **ESP** of the electronic structure solution and the classical nuclei, for short:

$$E_{\text{Coul}} = \sum_m q_m \underbrace{\langle \Phi^m \rangle}_{\text{ESP}}. \quad (2.72)$$

Such an **ESP** created by the molecule is a rigorously defined **QM** observable quantity^[278] and describes here the first-order interaction between a positive charge at any point, m , surrounding the **QM** system (which is in line with the perturbation treatment in Section 2.3.2).^[279] Calculating the **ESP** at different positions around the molecule would lead to the so-called molecular electrostatic potential. Higher-order effects are induction/polarization terms then, going up to infinite order, *i.e.*, including re-polarization. Full polarization of the **QM** part is included via the modified $\hat{H}_{\text{QM/MM}}$ (see Eq. (2.70) or Eq. (2.74) below).

2.4.3 Quantum Mechanics/Molecular Mechanics with Semi-Empirical Coupling

Within the usual **LCAO** description, this **ESP** in a **QM/MM** treatment becomes

$$\langle \Phi^m \rangle = - \sum_{\mu\nu} D_{\mu\nu} V_{\mu\nu}^m + \sum_J Z_J V^{mJ}, \quad (2.73)$$

with the electronic density matrix elements, $D_{\mu\nu}$ (with μ -th and ν -th atomic orbital basis functions). $V_{\mu\nu}^m = \langle \mu | \|\mathbf{r} - \mathbf{R}_m\|^{-1} | \nu \rangle$ describes the **GOCAT**-charge attraction integrals and V^{mJ} the Coulomb repulsion terms as the interaction of a probe unit test charge and an electron in the former and nucleus in the latter case.

As already described in Section 2.4.1, using Eq. (2.70) leads to the modification of the

³³Without all diverse interactions of a typical **FF**, adding Coulomb attractions or repulsions *between* the **GOCAT** charges themselves is not meaningful. The global optimization would then tweak the outer shell in order to create *any* energy that is possible within the feasible set and leads to a good fitness—without any physical content. If instead (all the) other interactions of an **FF** were included, we would not be “abstract” anymore but using a type of discrete and numerical optimization at the same time with at least “ghost” atom quality (atom types) behind each charge. This would lead to a *future* level of complexity of a **GOCAT**.

core Hamiltonian

$$\tilde{h}_{\mu\nu} = h_{\mu\nu} - \sum_m q_m V_{\mu\nu}^m. \quad (2.74)$$

The solution yields new perturbed, *i.e.*, polarized, density matrix elements, $\tilde{D}_{\mu\nu}$, and a new QM energy, accordingly. Using $\tilde{D}_{\mu\nu}$ in Eq. (2.73) leads to the interaction energy of MM and QM within the polarized (higher-order) picture: The relaxation of the density matrix due to the external point charge perturbation can be identified as polarization. More formally, the energy can be divided into the following contributions; for the induction/polarization energy part results

$$E_{\text{ind}} = \tilde{E}_{\text{QM}} - E_{\text{QM}} - E_{\text{Coul}}, \quad (2.75)$$

similar to Ref. [276]: \tilde{E}_{QM} is the energy after using Eq. (2.74) in an SCF, E_{QM} the energy without any coupling, *i.e.*, with no MM part present, and E_{Coul} is the (first-order) interaction energy between the point charges and the QM charge density, see Eq. (2.72).³⁴

For a semi-empirical level of theory, in contrast to *ab initio*, there are, however, some nuisances. As described by BAKOWIES and THIEL in Refs. [275, 276] and references therein,^[281–283] the semi-empirical treatment of the integrals as done in the NDDO approximation (*cf.* Section 2.2.3), leads to the neglect of some two-center and all three- and four-center integrals. The remaining integrals and the core-core repulsion are re-parametrized with suitable parametric formulas against experimental (and/or theoretical high-level) results. Two-center one-electron integrals as well as core-core repulsion integrals depend on these empirical parameters for both atoms, in semi-empirical theory. However, this should not be the case for $V_{\mu\nu}^m$ or V^{mJ} with respect to the “measurement” position of the ESP of the probe charge position m . Either the same integral approximations as used in the underlying semi-empirical level of theory can be made or an additional (complementary) re-parametrization to improve this deficiency, *i.e.*, to reach another reference ESP on a *ab initio* level.^[275]

Even further approximations, reaching a more classical-like picture, are investigated. In the seminal work of WARSHEL and LEVITT^[273] and subsequent studies,^[284] a coupling such as

$$\tilde{h}_{\nu\nu} = h_{\nu\nu} - \sum_m \frac{q_m}{R_{mJ}}. \quad (2.76)$$

was used, where the ν -th orbital is centered on atom J . That is to say, no non-diagonal Hamilton matrix elements are changed and the overall effect on the electronic part ($\sum_m q_m V_{\mu\nu}^m$ in Eq. (2.74)) is instead emulated by an additional ESP created at the atoms J . Intuitively stated, this will lead to an increase (or decrease) at the J th position to attract more (or less) electron

³⁴Polarization of the MM part would be the next logical step in a more realistic description of intermolecular interactions between QM and MM molecules, as a more balanced description. But having *per constructionem* just partial point charges in the GOCAT renders this impossible. For such a setting, (at least) dipoles and (their) polarizabilities would be needed on top of partial charges. This treatment could then be “equilibrated” with the QM part in an iterative fashion until convergence of mutual polarizations.^[274,276,277,280]

density within the molecule (so to speak a change of the atoms' electronegativities). This was implemented as part of a paradynamics-based free-energy calculation benchmark^[285] in MOPAC^[226] and used in big parts of this Thesis.

To put the last point into the perspective of this GOCAT endeavor, one can remark: The chosen level of theory for QM computations is always a compromise between still correctly treated physics and the computational efficiency needed for any *global* optimization, at least, if any chance of global convergence is to be reached. In this context, the introduced integral approximations of SQC also just emanate again at the QM/MM coupling level.³⁵

2.5 Minimum Energy Path

In most cases, for classical statistical theories, it is sufficient to find and relate the involved chemical elementary reaction steps and the energetic barriers between them in order to describe the kinetics of the chemical reaction system (*cf.* Chapter 1). The elementary steps are local energetic minima points on the PES that thus correspond to stable configurations. For the transitions between these, the connecting valleys that lead to the smallest energetic barrier between the minima are commonly the most important ones, *i.e.*, the rate-limiting paths. The highest energy point on such a connecting path is called the TS. It is a first-order saddle point on the PES. Therefore, the PES has a positive (convex) curvature in all directions except for one which has a negative (concave) curvature. In mathematical terms, the latter direction is that of the Hessian matrix eigenvector that corresponds to its single negative eigenvalue.^[142] If one follows this direction (or mode) at the TS along a steepest descent path down to the next connected local minima, this steepest descent path created is usually called interchangeably the MEP.^[287] In practice, one widespread scheme for finding a MEP is the *intrinsic reaction coordinate (IRC)* approach.^[287-291] In the *simplest* way, an IRC can be generated by following each positive and negative direction along the imaginary frequency mode downhill with complete quenching of the velocity, *i.e.*, with infinitesimal velocity, in an MD, using mass-weighted coordinates.³⁶ The MEP is associated with the most likely path a reaction takes and thus defines the most likely mechanism. Anywhere on this path, a tangent will be parallel to the nuclear gradient of the electronic energy. For illustration purposes, three minima with two saddle points (TSs) are shown in Fig. 2.5 on the next page. Therefore, the general goal is to locate exactly these connected stationary points on the PES such as the minima and the TSs.

Naturally, there can be multiple different paths connecting the same minima over different barriers or even further intermediate steps, *i.e.*, other local minima. Consequently, especially in more complex systems with many *degrees of freedom (DOFs)* and thus also many loose ones, this picture complicates seriously. In such cases, the PES can have several flat local minima with connecting low-energy barriers compared to the overall deep global minimum regime or funnel. In such cases, *one* MEP is not suitable to define the most likely mechanism,

³⁵ Interestingly, in other work within our group^[195,196] a local version of MOPAC for surface hopping dynamics was used with a different coupling.^[286]

³⁶ Without velocity quenching, one would have a so-called minimum dynamic path.^[292]

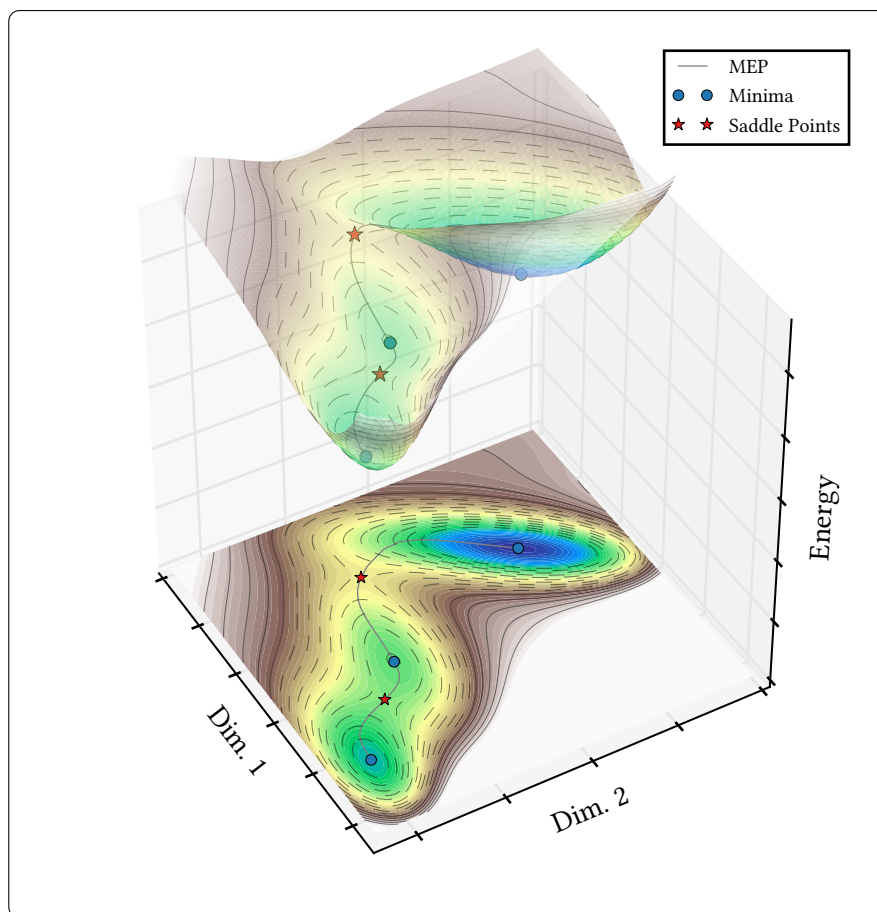


Fig. 2.5: PES illustration of arbitrary configurational coordinates, exemplified on the Müller-Brown potential.^[293] See the main text for further explanations.

simply because there might be even multiple equally important pathways connecting the stable minima. At finite temperatures and thus also incorporating entropic effects, a **free energy surface (FES)** together with reduced collective variable representations can average over such local features and thus include some thermal fluctuations. Often, a minimum-free-energy path can be defined after sampling a mean-force representation.^[294] Even this description, however, might fail for more complex systems.^[295]

As in the case of the present Thesis, for setting the scene of the **GOCAT** *ansatz*, we deliberately limited our investigations to small, already mostly well-understood systems, where such complexities are not yet crucial. Hence, for further discussions about this broad topic, we point to Refs. [295–297], including, *e.g.*, the branching of reaction channels and bifurcation points,^[295] comparisons between the **MEP** to a so-called minimum action path and their restrictions, as well as reviews about general methodologies on such complex **PES**, including, *e.g.*, reaction tubes and principal curves.^[296]

2.5.1 Nudged Elastic Band

There is a plethora of different methods for locating the important stationary points available in the literature. General optimization techniques for stationary points on a

PES are summarized in Ref. [142] and different (black-box) search protocols of reaction mechanisms were recently reviewed in Ref. [298]. Generally, one can divide those methods into three categories: (1) TS finding and IRC following, (2) single-ended and (3) double-ended methods. In (1), the TS must somehow be directly optimized,^[299] in the simplest case by choosing a good starting point in the neighborhood of the true TS and a subsequent TS optimization, with minimization in all directions but the one imaginary mode direction.^[142] A usual IRC connects both minima then. In (2), starting at one minimum a (hypothetical) reaction coordinate must be defined, and the procedure is started to reach the TS and finally a description of the MEP. In (3), both minima serve as starting points for the algorithm to optimize a MEP between those two. Specific other algorithms are reviewed in Ref. [298]. The method which was used and extended in this Thesis is the nudged elastic band (NEB),^[300–302] which belongs to the double-ended *chain of states*, or chain of frames, algorithms and will be discussed below. Maybe the strongest competitor also belonging to the chain of states algorithms is the *string* method that redistributes the frames after each cycle *via* interpolation techniques,^[303,304] its growing variant as a gradual build-up of the chain,^[305–307] and further successors such as the single-ended growing string method^[308] (compare with the already cited Ref. [298]).

In the case of GOCAT optimization as in the present work, usually two end-points of a reaction are known, *i.e.*, the R and P frames, as one wants to find embeddings to catalyze this specific mechanistic step. Then a MEP (more specifically, a steepest descent path)^[309–311] with a proper TS is reached with high reliability after full convergence of the NEB.³⁷ Especially in the “vertical” static GOCAT setting without relaxing the MEP, the discretization comes in handy as not just the saddle point, but the whole reaction path results. In this way, multi-modality of the path and further artifacts can be penalized, if they occur (compare with Section 3.6 where it is illustrated that using just a TS stabilization leads to overfitting). In the following, the version implemented for this Thesis is described, which was inspired^[317] by the similar functionality available in the program package ASE.^[318]

2.5.1.1 Improved Tangent and Climbing Image

First, define a *chain* of states—or synonymously: frames, images, beads—, $\{\mathbf{R}_0, \mathbf{R}_1, \dots, \mathbf{R}_N\}$, with $N - 1$ flexible frames, where \mathbf{R}_i are the (usually Cartesian) coordinates of *one* frame. In most cases, the first and last frame should be proper local minima, *i.e.*, the already locally optimized R and P frames:³⁸ These are fixed and not changed during the subsequent optimization of the other frames. NEB introduces a simple projection scheme for two different mutually orthogonal kinds of “forces”, \mathbf{F}_i , during the optimization (*cf.* Fig. 2.6 on

³⁷Procedures for TS finding for reaction mechanisms are generally not easily automated. For different methodologies using varying amounts of heuristics (and user-intervention) and also reaching reaction networks,^[312–316] including many different connecting paths between different intermediate, again Ref. [298] can serve as overview.

³⁸Though, use of, *e.g.*, the window-based sub-NEBs, *vide infra*, actually lessens the need of having strictly local optima at the two end-frames, as long as no kinks appear during optimization due to too strong gradients and/or erratic large optimization steps.

the following page):

$$\mathbf{F}_i^\perp = -\nabla E(\mathbf{R}_i) + \nabla E(\mathbf{R}_i) \cdot \hat{\boldsymbol{\tau}}_i \hat{\boldsymbol{\tau}}_i, \quad (2.77)$$

$$\mathbf{F}_i^{\text{S}\parallel} = (\|\mathbf{R}_{i+1} - \mathbf{R}_i\| k_i - \|\mathbf{R}_i - \mathbf{R}_{i-1}\| k_{i-1}) \hat{\boldsymbol{\tau}}_i. \quad (2.78)$$

So, orthogonal to the tangential direction to neighboring frames, $\hat{\boldsymbol{\tau}}_i$, each frame i (excluding the end-frames) is allowed to relax on the PES, *via* the force in Eq. (2.77). This force is the negative gradient of the PES where a PES contribution along the tangent direction is erased by the projection of the second term in Eq. (2.77). In parallel direction to the tangent in Eq. (2.78), an artificial force is introduced, which holds each neighboring frame apart to equidistant positions and prevents that frames slip down to the closest end-frame or closest local optima. If the Euclidean distances $\|\mathbf{R}_{i+1} - \mathbf{R}_i\|$ and the counterpart $\|\mathbf{R}_i - \mathbf{R}_{i-1}\|$ become unequal, neighboring frames begin to “nudge” each other in tangential direction, scaled *via* either frame-dependent or global spring force parameters k_i .³⁹ The final to be minimized force is simply the sum of these both

$$\mathbf{F}_i^{\text{NEB}} = \mathbf{F}_i^\perp + \mathbf{F}_i^{\text{S}\parallel}. \quad (2.79)$$

In the “upwind” scheme, the tangent is defined to be the vector to the neighbor of frame i with *higher* energy, which is the *improved tangent* version of NEB^[301]

$$\boldsymbol{\tau}_i = \begin{cases} \boldsymbol{\tau}_i^+ = \mathbf{R}_{i+1} - \mathbf{R}_i & \text{if } E_{i+1} > E_i > E_{i-1} \\ \boldsymbol{\tau}_i^- = \mathbf{R}_i - \mathbf{R}_{i-1} & \text{if } E_{i+1} < E_i < E_{i-1} \end{cases}, \quad (2.80)$$

and which is normalized then to $\hat{\boldsymbol{\tau}}_i = \boldsymbol{\tau}_i / \|\boldsymbol{\tau}_i\|$.

At (intermediate or final) energetic local optima during the NEB, a weighted mixture of both these cases is used

$$\boldsymbol{\tau}_i = \begin{cases} \boldsymbol{\tau}_i^+ \Delta E_i^{\text{max}} + \boldsymbol{\tau}_i^- \Delta E_i^{\text{min}} & \text{if } E_{i+1} > E_{i-1} \\ \boldsymbol{\tau}_i^+ \Delta E_i^{\text{min}} + \boldsymbol{\tau}_i^- \Delta E_i^{\text{max}} & \text{if } E_{i+1} < E_{i-1} \end{cases}, \quad (2.81)$$

with

$$\Delta E_i^{\text{max}} := \max(|E_{i+1} - E_i|, |E_{i-1} - E_i|), \quad (2.82)$$

$$\Delta E_i^{\text{min}} := \min(|E_{i+1} - E_i|, |E_{i-1} - E_i|), \quad (2.83)$$

as a switching function for a smoother tangent direction change during optimization.

This version of the tangent definition usually avoids kinks in flat area regions of the PES, *e.g.*, at the end-frames, and prevents corner-cutting. Otherwise, this would be problematic in curved MEP regions, where the higher energy/gradient regions are avoided and a frame is

³⁹Note that this most robust description—to the knowledge of the present author—uses magnitudes of the neighboring distances, which does not strictly force all frames to have exactly the same neighbor distance globally. But older schemes,^[300,301] and also the “improvement” in regard to this point mentioned in Refs. [319, 320] could lead to kinks at non-linear arcs of the MEP.

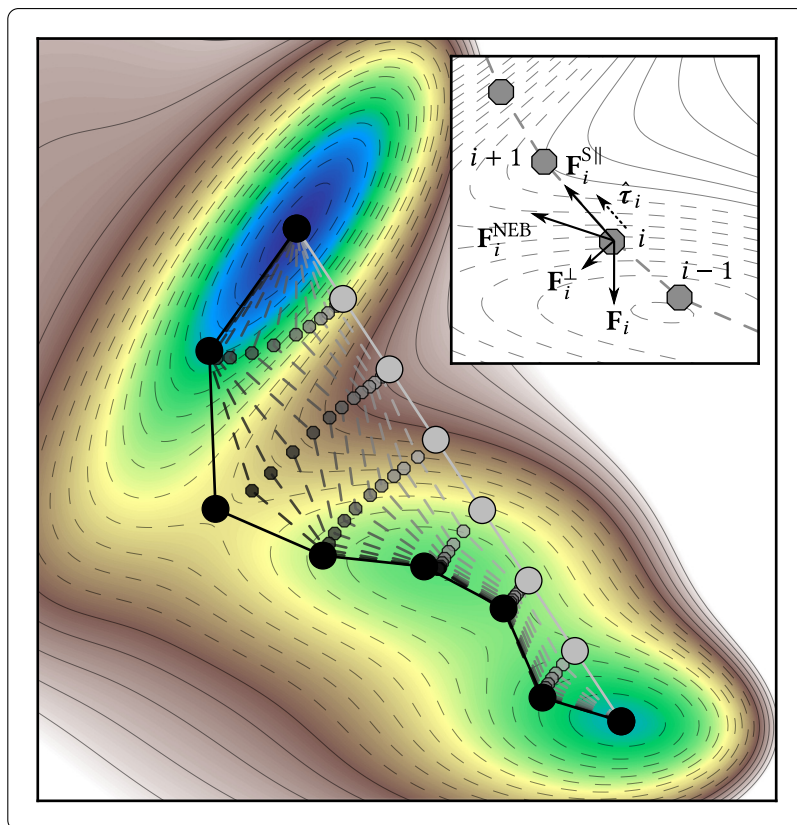


Fig. 2.6: NEB illustration (compare with Fig. 2.5): In the main Figure, different iterations of the NEB algorithm are shown, from the linearly interpolated starting path in gray to the finally optimized one in black. In the inset, one of these intermediate paths is shown again with the force vectors of Eqs. (2.77) to (2.79), using $\mathbf{F}_i = -\nabla E(\mathbf{R}_i)$ (force by the PES) and its projections due to the NEB formalism. Note that a simplest NEB implementation was used for illustration; no CI NEB was used, and the TS was, hence, not exactly found (see Eq. (2.84) on the following page). At the global minimum a “corner-cutting” artifact due to the discretization and local optimization dependent on the starting path without adaptivity can be stated. Improvements are discussed in the main text.

placed a bit more downhill.⁴⁰ The projection scheme of Eq. (2.79) is vital. Without it, as used in older *elastic band* versions, this would lead to a competition of forces of the PES vs. the band such that a chain of frames would be kicked off the MEP due to the band-based forces in the final iterations of the optimization.^[321] However, due to this projection scheme, the mixed second derivatives of the forces are not equal, *i.e.*, one cannot define a scalar objective function to be optimized. In other words, the forces are not conservative.^[321–323] Hence, just the forces (or gradients) should be minimized to zero by usual optimizers without sticking to a scalar objective function directly for line searches, convergence tests, etc. For this Thesis, several local optimization algorithms were examined, including also the very efficient LBFGS algorithm with and without line searches, restarts⁴¹ and other robustness

⁴⁰ Exceptions are rarely occurring numerical problems due to very loose DOFs or too steep gradients without using a step control (trust radius), especially at the beginning of the optimization.

⁴¹ Quasi-Newton algorithms that try to learn curvature of the objective function surface, but without having a symmetric Hessian in the NEB case, need restarts multiple times when the current gradient direction is very different from the proposed one influenced by the recent history of steps taken.

improvements.^[318] At the end, the **fast inertial relaxation engine (FIRE)**^[324] algorithm was used that was found to be the most robust optimizer.

After a first round of **NEB** optimization, the highest-energy frame is supposed to be just *near* a real **TS**. Thus, in a second round of **NEB**, a **climbing image (CI) NEB** can be started. Here, on the frame l that is the one of the highest energy, $E_l = E_{\max}$, another force is exerted than on all other frames

$$\mathbf{F}_l^{\text{CI}} = \mathbf{F}_l - 2\mathbf{F}_l \cdot \hat{\boldsymbol{\tau}}_l \hat{\boldsymbol{\tau}}_l. \quad (2.84)$$

The currently best estimate of the saddle point is hence allowed to climb uphill in energy into the tangential 1D direction of its neighbors while relaxing in all orthogonal directions.

Besides that, all external **DOFs**, *i.e.*, the 6(5) coordinates for translation and rotation, should be erased for calculating inter-frame, *i.e.*, inter-molecule, distances as long as **NEBs** are performed for gas phase paths. Otherwise, a finite distance between neighboring frames, which is build up and conserved due to the nudging, could be just a result of *identical* frames that are only shifted or rotated in absolute space. Without any “anchor” point such as, *e.g.*, the one within an external potential as that of an electrostatic **GOCAT**, the space homogeneity and isotropy^[186,325] should therefore be incorporated.⁴² This is done *via* optimal alignment, *i.e.*, center of mass transformations and optimal rotations based on quaternions using KEARSLEY’s algorithm.^[327,328] For **NEB within a GOCAT**, the *fixation* of the latter provides for such an anchor point.

2.5.1.2 Nonlinear Interpolation

For the initial path generation between the start and end frame, a non-linear interpolation scheme based on Ref. [329] is used. By this, collisions of atoms are avoided, which otherwise would lead to **SCF** convergence errors and thus to no gradients at those broken configurations at all. Additionally, the starting path is already closer to the final **MEP**, which saves some iterations for the **NEB** convergence. An artificial help potential, the **image dependent pair potential (IDPP)**, is defined, which motivates all internal atom-pair distances, d_{jk}^i , to be close to the target ones between atoms j and k for frame i . The trick is to use a *linear* interpolation of atom-pair distances (instead of Cartesian coordinates directly),⁴³ where α as superscript denotes the start frame and β the end frame

$$d_{jk}^i = d_{jk}^\alpha + i \left(d_{jk}^\beta - d_{jk}^\alpha \right) / N. \quad (2.85)$$

In the typical starting path, i starts at $0 = \alpha$ and ends at $N = \beta$ with having $N + 1$ total frames and $N - 1$ flexible frames. Note that N can also be a subset of these frames in the adaptive version of the **NEB** later on (*cf.* Section 2.5.1.3). These distances are then used in a

⁴²The **NEB** formalism was invented in the regime of surface chemistry, but also using Cartesian coordinates. Thus, overall artifacts by erratic translations and rotations were less a problem. For gas-phase paths, this problem was re-addressed by Ref. [326].

⁴³Others^[330,331] also use different (redundant) internal coordinates for **NEB** optimization which helps with the initial path generation, but Cartesian coordinates are supposed to be more stable.

sum of squared error objective function, which is the **IDPP**,

$$S_i^{\text{IDPP}}(\mathbf{R}_i) = \sum_j^{N_{\text{atoms}}} \sum_{k>j}^{N_{\text{atoms}}} w(d_{jk}^i) \left(d_{jk}^i - \|\mathbf{r}_j(\mathbf{R}_i) - \mathbf{r}_k(\mathbf{R}_i)\| \right)^2. \quad (2.86)$$

$\|\mathbf{r}_j(\mathbf{R}_i) - \mathbf{r}_k(\mathbf{R}_i)\|$ is the usual Euclidean norm between the *atomic* Cartesian coordinates, $\{\mathbf{r}_j, \mathbf{r}_k\}$, which are written as a function of a frame structure \mathbf{R}_i ; the term in parentheses is a summed squared error term, weighted by the matrix elements, denoted as function $w(d_{jk}^i)$.⁴⁴ Now, a **NEB** is performed on that **IDPP**, using $\mathbf{F}_i^\perp = -\nabla S_i^{\text{IDPP}}(\mathbf{R}_i)$ for Eq. (2.77) instead of the gradient of the **PES**.⁴⁵ For each interpolation that generates new frames, one round that consists of atom-distance interpolations and a subsequent **NEB** optimization on the **IDPP** is hence carried out. The resulting non-linear interpolated path is, in turn, the starting path for the **NEB** on the **PES**.

In Fig. 2.7, the different interpolation schemes are pictured. In Fig. 2.7(a) the molecular viewing program^[334] already detects “bonds” due to small radii, also between H-atoms, which is indicated here by the red color. Such structures will lead to high energies and gradients; if a collision appeared, which is not yet the case here, the **NEB** would be ill-defined.

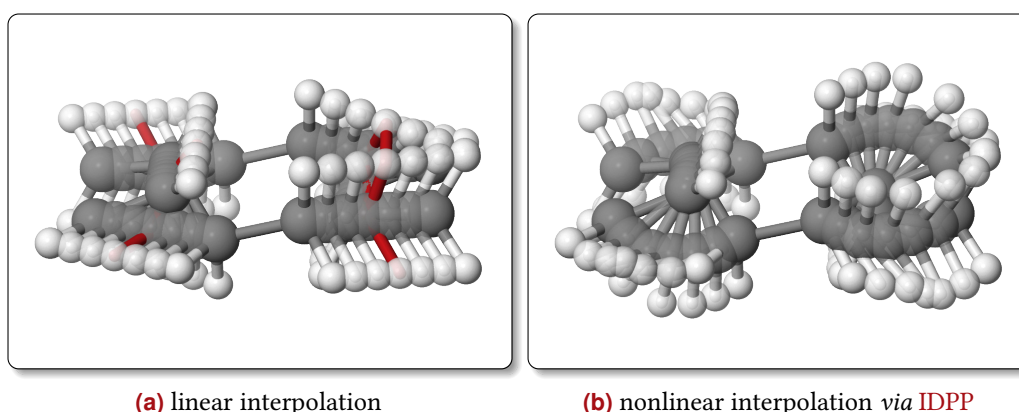


Fig. 2.7: For a simple Cope rearrangement of a 1,5-hexadiene, a linear and a nonlinear interpolation are shown, which are the starting paths for subsequent **NEBs** then. The red “bonds” drawn in Fig. (a) indicate a structure of smaller distances at some places, e.g., an “H-H” bond detected by the molecular viewer. Even “fusion” could happen by accident for some systems when a linear interpolation is used (not shown). Compare with Fig. 2.9 on p. 60 where the starting and end frames are shown, i.e., **R** and **P**. (Remark: This example is selected because it also serves as illustration for the adaptive **NEB** below, although there would have been systems with more apparent “fusions” of atoms at particular frames.)

2.5.1.3 Adaptive Nudged Elastic Band

KOLSBJERG *et al.* developed an *adaptive NEB* scheme,^[319,320] which is supposed to automate the **MEP** search. The optimized starting end-frames are still needed, but the total **MEP**

⁴⁴In practice $w(d_{jk}^i) = (d_{jk}^i)^{-4}$ was used in order to place more emphasis on smaller atomic distances.^[329]

⁴⁵Without using a **NEB**, it could lead to discontinuous interpolated paths as known from the usual linear synchronous transit approach,^[332] which is otherwise very similar to **IDPP**.^[329,333]

optimization, which consists of multiple rounds of adaptive interpolations/NEBs and a final CI in this variant, is made more robust and carried out all in one algorithm. This extension to the NEB fulfills the following purposes here:

- Not having to optimize all $N - 1$ frames at once, but just a window of $N_{\text{sub}} < N - 1$ frames. This division into sub-NEBs⁴⁶ will save multiple SP calculations.
- Adding frames gradually until $N - 1$ frames (excluding the fixed endings) are reached which incorporates a protocol based on an energy-to-geometry ratio, ζ (*vide infra*).

Firstly, a rough path is optimized, followed by a piece-wise optimization of paths centered around newly added frames, including some neighboring frames, where at each sub-NEB the addition of frames *focuses* on the important regions. The latter point is of crucial importance because for highly varying energetic barriers or high geometric curvature of the final MEP, the interpolated initial-guess of the NEB might be far off the final MEP (for instance, compare Fig. 1 of Ref. [319]). This can hardly be known *a priori*, and hence *gradually* adding frames where they are *currently* needed remedies this deficiency.

The protocol for frame-addition is specified as

$$\frac{f(\{\mathbf{R}_0, \dots, \mathbf{R}_N\})}{g(\{E_0, \dots, E_N\})} > \zeta, \quad (2.87)$$

$$f(\{\mathbf{R}_0, \dots, \mathbf{R}_N\}) = \frac{\max(\{|\mathbf{R}_1 - \mathbf{R}_0|, \dots, |\mathbf{R}_N - \mathbf{R}_{N-1}|\})}{|\mathbf{R}_N - \mathbf{R}_0|}, \quad (2.88)$$

$$g(\{E_0, \dots, E_N\}) = \frac{\max(\{\dots, \overline{\Delta E}_i, \dots\})}{E_{\text{norm}}}. \quad (2.89)$$

ζ is a user-defined meta-parameter as ratio where to add a new frame adaptively. Eq. (2.88) is the currently maximal frame-neighbor distance of the NEB path, $\{\mathbf{R}_i\}$, normalized by the total (starting) distance between the end-frames N and 0 . Eq. (2.89) refers to the highest energy difference. Therefore, if the inequality of Eq. (2.87) is true, the geometrical gap is chosen and the energetic gap otherwise. Accordingly, for $\zeta = 0$, a new frame would always be added into the currently biggest geometrical gap; for $\zeta \rightarrow \infty$, the biggest energetic gap, *i.e.*, regions of rapid change of the PES, would get a new frame.^{[320]47} The further energy and geometry definitions of Eqs. (2.88) and (2.89) are

$$\overline{\Delta E}_i = \Delta E_i \cdot \frac{E_{\text{mean},i}}{E_{\text{norm}}}, \quad (2.90)$$

with

$$E_{\text{norm}} = E_{\text{max}} - E_{\text{min}}, \quad E_{\text{mean},i} = 0.5(E_{i+1} + E_i - 2E_{\text{min}}), \quad \Delta E_i = |E_{i+1} - E_i|. \quad (2.91)$$

These are shown in Fig. 2.8 on the following page. For each frame addition, the neighbor NEB spring constants, k_i , are changed accordingly in order not to introduce an initial spring force in the next sub-NEB iteration.

⁴⁶The idea of adaptive sub-NEBs were proposed already earlier in another protocol in Ref. [335].

⁴⁷Mostly, the authors' recommendation in Ref. [319] of $\zeta \approx 0.5$ was employed.

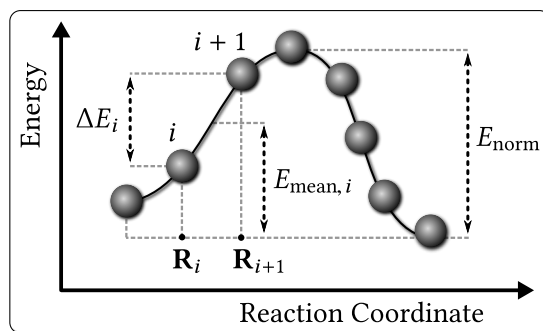


Fig. 2.8: Illustration for the definitions for Eqs. (2.90) and (2.91). Adapted from Ref. [319].

A typical final result of an adaptive NEB for the Cope rearrangement (*cf.* Fig. 2.7 on p. 57) is illustrated in Fig. 2.9 on the following page. The final gradient norms at the stationary points are very small, as expected $\|\nabla E\|_{\{R, TS, P\}} < 0.25 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, see Fig. 2.9(a). Here, the high curvature with regard to the energy and also some curvature along the geometric coordinate would be quite challenging for a non-adaptive NEB since the TS is peaked around 6–8 Å on the reaction coordinate, while larger parts along the reaction coordinate are less energetically varying. The coordinate itself is defined here as discrete Euclidean distance between all the frames, starting at 0 for the first frame, **R**. In this region around the **TS**, new frames were added adaptively and increased the resolution there. For this reason, the **CI** was then enabled to find the correct **TS** along the 1D coordinate dictated by its neighbors. If the latter 1D coordinate had not included the exact real **TS** and had not so after many further relaxing NEB iterations, the correct **TS** could not have been found.

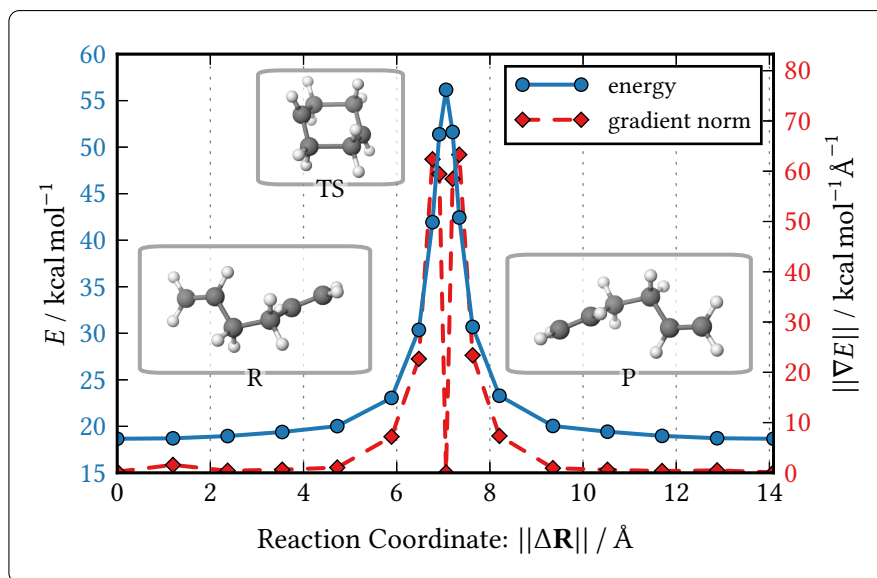
In Fig. 2.9(b), the corresponding final 19 frames of the **MEP** are given. These final converged frames, clearly, are already more similar to the non-linearly interpolated ones as starting path that was illustrated in Fig. 2.7(b) on p. 57 than the linearly interpolated ones of Fig. 2.7(a).

2.5.1.4 Possible Improvements

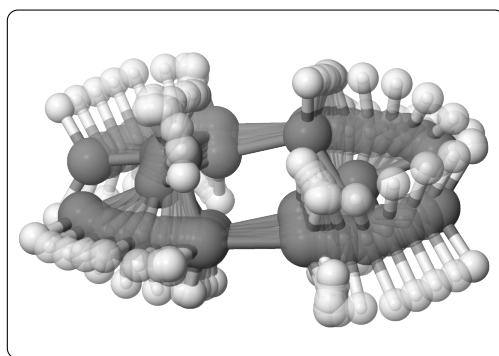
There are numerous further studies with respect to the NEB formalism and modifications thereof, including specific applications.⁴⁸ In the future, some of these could also be fruitful for further improvements of both the robustness, *i.e.*, a convergence without artifacts such as not finding a proper **TS**, as well as efficiency. These variations include, *e.g.*, so-called double-nudging^[337] for increased performance at the beginning of the optimization or for long paths.⁴⁹ Then, a switching function for smooth progression between single-, which is the usual nudging in NEB above, and double-nudging should be used.^[321] Related to this in other studies, also some perpendicular forces from the band itself have been mixed into the total force based on switching functions concerning the angles between each frame and its

⁴⁸ This small Section considers only improvements on NEB-based approaches; for completely different MEP optimization algorithms, see Section 2.5.1 and Ref. [298].

⁴⁹ Actually, this variant as well as some of Ref. [319] were implemented for this Thesis but must be benchmarked and should be extended to include proper switching functions for the former.



(a) energies and gradient norms after adaptive NEB



(b) superposition of all 19 frames

Fig. 2.9: Example of the final result of the adaptive NEB (on PM6^[214] level of theory). This reaction besides many others of a benchmark set in Ref. [336] was used for testing all the implementations of this Section. Compare with the non-linearly interpolated starting path of Fig. 2.7(b) on p. 57.

neighbors.^[300,338] With a standard scalar product angle between three frames defined as

$$\cos \theta_i = \frac{(\mathbf{R}_{i+1} - \mathbf{R}_i)(\mathbf{R}_i - \mathbf{R}_{i-1})}{\|\mathbf{R}_{i+1} - \mathbf{R}_i\| \|\mathbf{R}_i - \mathbf{R}_{i-1}\|}, \quad (2.92)$$

each angle should be zero in the limit of *infinite* frames, independent of the actual curvature of the path. Using a finite number in practice, this can be used as indicator for undesirable kinks or wrong convergence behavior.

Moreover, an advanced interpolation scheme reformulated as a search for the geodesic curve on a Riemannian manifold by using proper metrics directly in the Cartesian domain was presented.^[333] This could also be used as initial guess for a NEB optimization instead of the IDPP-based one. Furthermore, spectator modes and special definitions for the distance calculations between the frames suited for complex environments (enzymes) with soft DOFs were discussed.^[338] Even NEB on free energy surfaces by frame-wise umbrella integration^[339] was developed.^[340] Regarding the optimization algorithms, protocols for

super-linear progression based on Newton-Raphson variants^[322] were presented, LBFGS variations^[321,330] and quite recently, a quadratically converging Newton-Raphson scheme using the full (non-symmetric) Jacobian matrices for finding the roots of the gradient/forces instead of the extrema of a (non-existent) scalar-valued function.^[323]⁵⁰ Besides, multiple CIs during NEB were proposed,^[341] a symmetry-aware NEB^[342] and further acceleration techniques were put forward such as, e.g., ANN potentials^[343] fitted to training data consisting of the NEB SPs or by using gaussian process regression (GPR).^[344] The latter acceleration methods try to bypass costly QM calculations by generating a ML potential on the fly.

In fact, this leads us to the next topic of this Chapter which covers some background of ML-based analysis that was used in the present work. Since the final pool of solutions of the GOCAT designs are often of high variance, especially unsupervised learning approaches in ML are beneficial for further analyses and discussions.

2.6 Machine Learning

Undeniably, ML has become virulent in the recent years in the domains of science and engineering—just think about the digital transformation happening nowadays affecting also industries, organizations and society as well. Besides other developments, this also includes all types of structure inferences from (big) data and thus leads to machines that learn to, e.g.,^[345] recognize objects, fingerprints, faces, to process speech and natural language, to control robots, to drive cars,^[346] and to machines that can play sophisticated games such as chess or Go.^[347] With the increase of computational resources as well as the inferential/statistical algorithms on the one hand and the data and their interconnections available on the other hand, this has even lead to the concept of “data-driven” science^[348] (for further thoughts on this topic, including some philosophical ones, the (German) perspective in Ref. [349] can be consulted.) As a matter of course, such developments towards this data-centric, (statistically) inductive, empirical approach also permeates theoretical and computational chemistry nowadays, with a continuously growing application spectrum.^[350,351]

The umbrella term ML can be seen as sub-field in AI which encompasses all types of algorithms and statistical models that improve or learn “automatically” through experience or training (besides the general vagueness of these terms, this refers to the term creation by MITCHELL).^[352] The aim is to solve problems without *explicit*, specifically programmed instructions for the respective problem but by relying on *general* patterns and statistical

⁵⁰Though, at least in the current state, analytical Hessians of the PES of the systems studied were required which impedes the utilization of this algorithm in more practical applications. Notwithstanding, since the Hessian could be calculated as by-product of the MDs in umbrella sampling, it was also used in Ref. [340].

inferences from data instead.^[353–355]⁵¹ The machine then is able to execute tasks that were not explicitly defined in the code beforehand.

Recent progress of **ML** in chemistry were reviewed multiple times: A more general view in chemistry can be found in Refs. [351, 357]. For atomistic (**QM**) property learning approaches of bypassing the Schrödinger equation, numerous routes were put forward for atomistic simulations and electronic property predictions.^[56,58,358,359] More focused on finding new materials, reviews and perspectives how to leverage **ML** in the domain of cheminformatics and materials science were also recently given,^[57,360–362] and with a focus on *deep learning* in Ref. [363]. To mention just a few recent examples at the boundary between global optimization and **ML**, there were **GAs** applied to the optimization of the training set for **ML**^[364] and **ML** techniques for niching^[365] as well as the **ML**-estimation of optimal balance between exploration vs. exploitation.^[366]

For the present work, actually only a small subset of **ML** algorithms finally accomplished to appear in this Thesis. Thus, there will be no extensive overview of the plethora of applications in chemistry and methods of **ML** given here; instead the cited literature should be inquired. Only some background of the used algorithms in the further Sections (Chapters 4 to 7) will be briefly described in the following.

Supervised learning: **ML** is usually categorized in *supervised learning* and *unsupervised learning*, as well as mixed versions thereof (semi-supervised). In the first case, the supervised learning, there are N pairs of sample/example data, also called *instances*, of $\{x_i, y_i\}_i^N$ and the task is to learn a function $f: x \mapsto y$.⁵² After learning from such input pairs, which *has* to include some generalization to yet unseen input,⁵³ the task is to predict \tilde{y} based on a new \tilde{x} . If there is a continuous output domain for y , this is called *regression*, and if it is categorical, *i.e.*, a finite set of output possibilities, this is a *classification* problem instead. The separate input values x of a sample are usually called *features*. Note that this also subsumes all types of (non-)linear ordinary multiple regressions, but of course also many other methods with very specific forms of f that defines the specific **ML** algorithm: Typical often used ones are **GPR**, support vector machines, **ANNs**, (kernelized) regressions, random forests, to mention just a few. The “learning” itself is often based on an optimization (of weights/parameters) with respect to the *training set* (of the N samples) with regard to the (training) error of predicted and known $\{y_i\}$ for fitting the specific **ML** model function f . For instance, this could be the mean-squared error (MSE), $\text{MSE} = 1/N \sum_i^N (y_i - f(x))^2$, or analogue measures; there are, however, also other methods without such fitting process as, *e.g.*, just by instance-

⁵¹In this context, also the other buzzwords such as **AI**, **ML**, big data, data mining, pattern recognition, etc., can be mentioned. These terms were created based on slightly different emphases, having evolved from different perspectives, such as focusing on the application of emulating something like an intelligent machine,^[356] or focusing on mere structure detection, while tackling problems for science or more so for engineering, etc., or other historical origins with somehow differing focuses. At the end, there is a huge *intersection* of algorithmic grounding. In this Thesis, only the term **ML** will be used.

⁵²This is given here in the most typical case with one scalar output and multiple scalar input variables. Of course, the data representation is another very important topic and there are also methods for arbitrary input and output domains.

⁵³Just memorizing the data would degrade the machine to a simple lookup table which is not what is meant by *learning*.

based regression using only a similarity metric (in simplest form: (k) nearest neighbor regression). The out-of-sample error or *test* error is the prediction error of not yet seen data which is not used for the training phase and leads to the topic of (under-/) *overfitting*. If the training error is low, but the test error is not, apparently the model does not *generalize* well and thus this would be dubbed overfit. Generally speaking, using a more complex model such as a (complex) deep ANN instead of just a simple linear regression for example,⁵⁴ this would impose a *lower bias* and a *higher variance* of the model and is more prone to being overfit. Then not only the actual true structure of the data is followed or learned, but also the actual, intrinsic (irreducible) errors or (erratic) noise.^[355] At the end, ML for regression is an arbitrary flexible (and diverse) “interpolation” approach between and based upon the known data, without the expected possibility of extrapolation in unknown regimes. Put differently, the generalization error *must* decrease with increase of training set size as long as this is machine *learning*. Usually, the representation (feature selection), the model complexity (including which family of algorithms to use and which meta-parameters of the models that also often control the bias-variance-tradeoff), the training set size, additional regularization terms to increase bias again, etc., must be considered to find an optimal ML model for the specific tackled problem. This last point can be considered as being most important for meaningful applications, especially as ML algorithms are routinely used also by non-experts due to freely available libraries and frameworks (e.g., Refs. [367–369]).

Unsupervised learning: In the second, the unsupervised approach, only the input $\{ \mathbf{x}_i \}$ is given and the task is to find structure in the data. This usually subsumes *clustering* (or cluster analysis) methods and *dimensionality reduction* approaches. In clustering, the aim is to find a grouping, *i.e.*, the cluster, of similar samples, which in turn are more dissimilar to the samples of other found groups, or equivalently, finding clusters as more dense regions that are separated from less dense regions in the feature space.^[370,371] In dimensionality reduction, the goal is to find a representation of the data in a lower-dimensional sub-space (or also possible in an arbitrarily non-linear manifold) in order to investigate and visualize the most important *relations* of the usual high- (or infinite-)dimensional data.

Force fields in the ML context: Note that the global optimization of REAXFF (*cf.* Section 2.2.4.1), as later shown in Chapter 4, is, indeed, a regression problem. As already pointed out by BEHLER in the context of his highly recognized **high-dimensional neural network potentials (HDNNPs)** for atomistic simulations,^[372–376] ML-based potentials can be seen as “mathematical” ones: Here an arbitrary function, f , without any physical basis by the ML model itself is introduced, as done in BEHLER’s ANN with specific feature representations and architectures, including specific descriptors, *i.e.*, the so-called *atomic symmetry functions*.^[377] Conversely, there are the empirical potentials, the FF, for atomistic simulations that directly are oriented along physically meaningful ingredients, *i.e.*, the decomposition of energy in low-dimensional (internal) coordinate-based terms, usually. Here,

⁵⁴Note that the hyperparameters and other model decisions about the complexity of the actual method used are also very important in this regard. But this again lies outside the scope of this Thesis and is well described in the usual textbooks.^[353–355]

potentials and thus parts of f encode the electrostatics (Coulomb terms), **vdW** interactions (e.g., **LJ** potential), and of course bonding terms in the most simplistic forms. However, with the raise of the complexity of the **FF**, including reactivity, multiple more terms, bond-order dependent terms, correction terms, etc. (cf. Section 2.2.4.1), there is a growing portion of “mathematical flavor” incorporated. In other words, the *physically motivated* terms, i.e., the bias, of an **FF** already determines that atoms and molecules can be simulated and that also the generalizability is usually better *per constructionem*. Yet, by the introduction of the multiple further non-empirical (correction) terms in REAXFF, this **FF** approaches the regime of “non-physically motivated” potentials. This is exactly the reason why global optimization is actually needed in the first place, but also why problems such as overfitting are an important topic.

2.6.1 Multidimensional Scaling

Multidimensional scaling (MDS) is one class of algorithms for non-linear dimensionality reduction in unsupervised learning which was used in this Thesis and is briefly described here, besides many other techniques with different (dis-)advantages.^[353,354,378,379] It was used as more *intuitive* approach, mainly for data visualization in 2D or 3D, to illustrate the *similarities* between the data points. The goal is very straightforward: Find a (locally) optimal representation of the raw similarities or distances in the original (high-) dimensional space, using Euclidean distances in the new lower (usually 2D) abstract Cartesian space, which can be rotated freely because this does not change the distances. The objective can, for instance, be expressed as a least-squares problem in the *metric MDS*,⁵⁵ called the *stress*

$$s(\{\mathbf{z}_n\}_n^N) = \sum_{i,j \neq i} (d_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|)^2. \quad (2.93)$$

This can work on *arbitrary* distances $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ between the N instances, $\mathbf{x}_i \in \mathbb{R}^p$, and map those to the new Cartesian space $\mathbf{z}_i \in \mathbb{R}^k$ with $k \ll p$. Note that there are more efficient optimization strategies than, e.g., gradient descent, for such an **MDS** available, using an iterative approach starting at random \mathbf{z}_i and using the concept of *majorization* (here, finding the minimum by iterative minimization of surrogate functions and some other algebraic “tricks”).⁵⁶ This brings in a better convergence rate and monotonic behavior, but can lead to different (locally optimal) results.

⁵⁵ There are also other variants and objective functions definitions, e.g., only working on the data on an *ordinal* scale, and the *classical MDS* working directly on Euclidean distances between the original points; in this case, it can deliver equivalent results to **principal component analysis (PCA)** as another (well-known) linear dimensionality reduction technique. The latter was also rarely used for analysis of the variance of the data along the main components, as, e.g., in Section 6.2; cf. the textbooks^[353,354] for more details.

⁵⁶ This leads to the so-called **scaling by majorizing a complicated function (SMACOF)** algorithm for **MDS**, proposed originally in Ref. [380], and well described again in Ref. [381]. This is also the algorithm used for the **MDS** in **SCIKIT-LEARN**^[367] and this Thesis.

2.6.2 Hierarchical Clustering

Hierarchical clustering (HC) is again *one* (of many) approaches^[370,371] for unsupervised learning utilized for the cluster analysis part: Here, also a distance matrix, D , must be encoded that measures the (dis-)similarity between the instances, similarly to **MDS** above. In the *agglomerative* or *bottom-up* approach of **HC**, each instance first of all defines its own cluster on the lowest level. Then, the distance between the instances as well as the computed lowest intergroup distances that are calculated from the former by different methods are used to merge clusters from the lower level to generate clusters on a higher level. This forms the hierarchy, as the name suggests. At the end, $n - 1$ levels are created with *one* cluster at the highest level containing all the data. This can also be illustrated as a rooted binary tree that is called the *dendrogram*. An illustration is given in Fig. 2.10. Note that the abscissa has no meaning in this dendrogram, whereas the ordinate shows the (computed merge) distance which is based on the original distances and the linkage function.

There are different *linkage* methods available to compute the intergroup distances between the sets G and H , as for instance (as corner cases):

$$d_{\text{single}}(G, H) = \min_{i \in G, j \in H} d_{ij}, \quad (2.94)$$

$$d_{\text{complete}}(G, H) = \max_{i \in G, j \in H} d_{ij}, \quad (2.95)$$

$$d_{\text{average}}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{ij}, \quad (2.96)$$

where $N_{\{G, H\}}$ denotes the number of instances of that group. The “average” linkage method in Eq. (2.96) is more precisely called **unweighted pair group method with arithmetic mean (UPGMA)** and was often used in this Thesis, as this best respects or preserves the original similarities, d_{ij} , during the cluster merging by treating all these on an equal footing since also the size of each cluster is taken into account. This was also frequently checked with the so-called *cophenetic correlation* between the raw similarities and the linkage similarities.^[354]

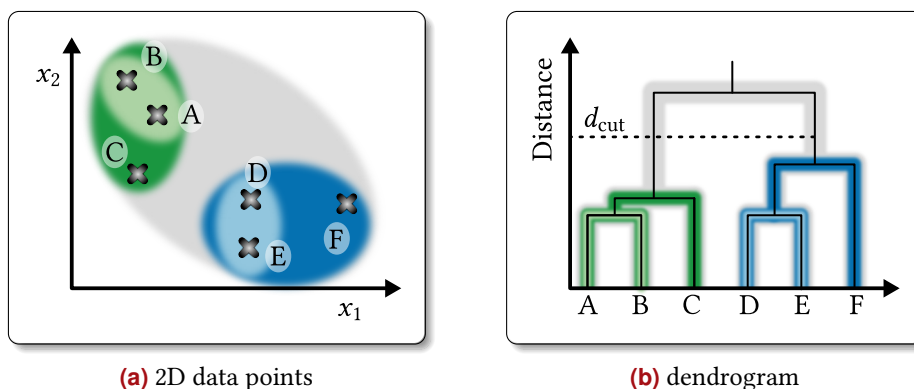


Fig. 2.10: An arbitrary data set, here already in two dimensions, is shown in Fig. (a). The data set is hierarchically clustered and the result is illustrated *via* a dendrogram in Fig. (b). The binary tree can then be intersected at an arbitrary value, d_{cut} , to generate an intended number of clusters.

Hence, as compromise between the other linkage methods, this tends to produce relatively compact clusters that are relatively far apart.^[354] Eq. (2.94), on the other hand, could lead to a *chaining* problem as only one (smaller) distance can lead to a merge of the clusters, irrespective of all other distances. In the end, **HC** is deterministic, but the final results can be very dependent on the data and possible outliers. This is also true for Eq. (2.95), but by enforcing maximal distances, this usually leads to spread, small clusters.

Without having to specify *a priori* how many clusters there should be, as it would be done, *e.g.*, in *k*-means clustering, there are also different variants available for clustering: Simply cutting the dendrogram somewhere horizontally at a certain height can lead to an intended number of clusters. This is shown in Fig. 2.10 on the previous page by d_{cut} , leading to two clusters in this constructed example. Besides, *e.g.*, there are *inconsistency* checks available that relate a merge/fusion height of the cluster tree to the heights below that, divided by their standard deviation, which tries to find a “natural” division into clusters. A sudden difference in subsequent linking heights points to an unlike distance between the clusters *vs.* the distances of the instances they contain.

As shown in the later Chapters 6 and 7, there are very many (and similar) realizations of the (electrostatic) **GOCATs**, such that a cutting criterion was mainly chosen by visual inspection of the results, which was also supported by the maximal acceleration of the cluster distance growth (*i.e.*, based on the discrete second derivatives, the so-called *elbow*-criterion). These clusters are then used as summarization of the results, *i.e.*, for finding best representatives in the search space, and also for choosing variable different starting individuals for, *e.g.*, the translations between the levels of theory (discussed later in Section 6.3.2). Note also that many of the implemented “dynamic” niching techniques are similar to a single linkage strategy in practice, as described in Chapter 3 and used for the **CSO** application in Chapter 5.

2.6.3 Distance Metric

Apparently, the most important (and sole) further input ingredient for the aforementioned techniques is the definition of a distance between data points. This leads also to the general problem of *representation*, *i.e.*, how to define chemically meaningful features.

The most important general requirements the descriptors for such **ML** usually must fulfill are first of all the important physical invariances under translation, rotation and permutation of any (indistinguishable) atom. For instance, using simply plain Cartesian coordinates and a distance such as the **root-mean-square deviation of atomic positions (RMSD)** between two systems/molecules, i, j :

$$d_{\text{RMSD}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\frac{1}{N_{\text{atoms}}} \sum_n^{N_{\text{atoms}}} \|\mathbf{x}_{i,n} - \mathbf{x}_{j,n}\|_2^2}, \quad (2.97)$$

encoding pairwise distances between the coordinates, this would also show translational and rotational symmetry, if aligned structures were used (see similar alignment for the **NEB** in Section 2.5.1.3). However, this would show no permutational symmetry: The exchange of

only two Cartesian vector elements of, e.g., x_i , would (probably) lead to $d_{\text{RMSD}} > 0$ even if $x_i = x_j$ holds *before* swapping these Cartesian elements, i.e., meaning the exchange of alike, actually indistinguishable atoms. Without being able to try every possible permutation of both included molecules to find the best (alike) order, other approaches must be considered. This translates also to representations and metrics for comparing clusters as in **CSO**.

Regarding *general* approaches, prominent examples for **HDNNP** are the already mentioned atom-centered symmetry functions (carrying this name because of these incorporated physical symmetries).^[375,377] Another quite often used representation is the **smooth overlap of atomic positions (SOAP)** method.^[382] As pointed out in recent studies,^[378,383] such *local* descriptors have in common to use atom-centered functions, e.g., Gaussians or, as a limiting case, delta-distributions, for describing the atom-density, which makes them permutationally invariant from the beginning. For translational and rotational symmetry, projections onto other functions can take place such as onto the symmetry functions. These depend on bonds and angles, etc., for incorporating two- and three-body interactions. Instead, an explicit symmetrization of a *kernel* function can be carried out as done, e.g., for **SOAP**^[378,383] (for further information of the implicit mapping using kernels in **ML** methods, see Ref. [358]).

Moreover, there are more *global* descriptors, as the **coulomb matrix (CM)** descriptor^[384] and its modification to **bag of bonds (BOB)**^[385] or, including higher-order interaction terms (three-body, ...), into the “BA-representation” (for bonds, angles, torsions, similar to usual internal **FF**-based representations).^[386] The latter representation was also introduced for the *uniqueness* property:^[387]⁵⁷ Using just pair-wise interactions or distances encoded into the descriptors can lead to no differentiation of *homometric* molecules. These are actually different molecules, but they do show the exact same distances between all involved atoms. This situation is actually also something that should be considered if using similar descriptors for clusters, especially when using just one atom-type or the indistinguishable entities of a **GOCAT** (point charges). Actually, the list and current research of possible descriptors and, as a result, also the definition of features for the **ML** in chemistry is growing quickly, since this is literally the *basis* of **ML** in chemistry and is needed especially for very accurate property predictions and **ML** potentials.

As in the current Thesis the descriptors were used for the unsupervised setting and the goal was thus to summarize results (find representatives) and build hypotheses, the requirements on descriptors are less severe and, furthermore, cannot be benchmarked by using an error of prediction. Thus, further discussions about such representation problems are outside the scope of this Thesis, and again we point to the relevant literature (and references therein).^[358,378,382,383,387,388] Note that there also is a plethora of representations/descriptors available that are *problem-specific*. In many applications, also in the regime of cheminformatics, often only very specific types of information or molecular characteristics have to be encoded for the target (e.g., **quantitative structure-activity relationship (QSAR)**/**quantitative structure-property relationship (QSPR)** or virtual screening approaches).^[159] Similarly,

⁵⁷ For these *ab initio* **ML** representations for property predictions, besides the mentioned demands, furthermore *continuity*, ideally *differentiability*, *generality* (encoding arbitrary chemical systems, including periodical ones) and *efficiency* with regard to computational cost and with regard to the prediction power are required.^[388]

there are (very many) specific representations and heuristics used for **CSO**, some of which are reviewed in the application in Chapter 5: For niching, distance measures based on the aforementioned **CM** variants and more problem-specific atomic neighborhood lists/binnings were used (compare also with the depictions in Section 3.5.3).

For **GOCAT** design and its analysis by **HC** and **MDS**, mainly symmetrized versions of **ESP** vectors are used (taking into account the overall symmetry of the reaction):

$$\varphi_J^{\text{ESP}} = \sum_i^{N_{\text{Ch}}} \frac{q_i}{\|\mathbf{r}_i - \mathbf{R}_J\|_2}, \quad (2.98)$$

i.e., the **ESP** generated by the charges, $\{q_i\}$, at the J -th core atom of, usually, the stationary structures. These are the **ESP** vectors, $\boldsymbol{\varphi}^{\text{ESP}} = (\varphi_1^{\text{ESP}}, \varphi_2^{\text{ESP}}, \dots, \varphi_{N_{\text{atoms}}}^{\text{ESP}})^T \in \mathbb{R}^{N_{\text{atoms}}}$, as discrete representatives of the **ESP** within a **GOCAT**. Then the $p \in \{1, 2\}$ norm is used to induce the metric and to calculate distances between two such reaction path **ESP** vectors

$$d_{ij} = d(\boldsymbol{\varphi}_i^{\text{ESP}}, \boldsymbol{\varphi}_j^{\text{ESP}}) = \left(\sum_n^{N_{\text{atoms}}} |\varphi_{i,n}^{\text{ESP}} - \varphi_{j,n}^{\text{ESP}}|^p \right)^{1/p}. \quad (2.99)$$

Besides that, a *duplicate detection* is implemented *via* the **BOB** descriptor for the outer surrounding, in combination with this **ESP** descriptor. This is needed to delete redundancies in the database after multiple (separate) **GA** runs. Both the **CM** and **BOB** descriptors are not explained *here* again, but they are introduced later in Chapter 5 and Section 6.2, including the duplicate detection. This detection simply erases candidate solution entries having distances below a certain threshold for both descriptors.

Methodology and Implementations

This Section sheds light onto the actual methods used and implementations needed for the present work. First of all, the used and extended program package OGOLEM is introduced in Section 3.1 that also reviews all applications done so far with OGOLEM in our work group. Next, an overview on the general extent and the state of the code base is given in Section 3.2. In Section 3.3, OGOLEM's main optimization intrinsics are exposed. Following in Section 3.4, a rough sketch of the new implemented features are given, some of which are illustrated in detail in Section 3.5. Finally, the most important ingredient, the chemically meaningful encoding of *catalysis* in the fitness function is discussed in Section 3.6.

3.1 OGOLEM

OGOLEM is a global optimization framework leveraging (mainly) the GA-based metaheuristic for nondeterministic optimization in different chemical contexts (*cf.* Section 2.1.2.2). It was started by DIETERICH as PhD project^[389] in our work group in 2010, based on older in-house developments we had until then, and is developed further ever since. Already right from the beginning, mainly three different types of optimizations were targeted: CSO, parameter optimization and *discrete* molecular design.

In the following, a concise overview of certain applications are illustrated as they were started with the seminal papers of DIETERICH in our work group and extended afterwards by him and his colleagues at different places. Parallel to this, also additional research conducted with OGOLEM as well as further developments of the package carried out in our work group are exemplified, excluding applications by other work groups.

Applications: After the initial framework paper,^[390] the triad of contexts was introduced by optimizing strongly mixed LJ clusters—with multiple different species of rare gas atoms in one cluster—and this work also already included a global parameter optimization of the vdW empirical potential of LJ-type with respect to high-level *ab initio* data.^[391] Following

this track, a Gupta potential^[392] was optimized against high-level reference data for mixed alkaline earth metal clusters that were optimized subsequently. Bigger flexible molecules (Kanamycin A) around physiological cations were tackled, which also included the need to optimize *internal* degrees of freedom in a joint theoretical and experimental study.^[393] Photoisomerizable molecules with varying substituent patterns in form of capto-dative groups on an azobenzene scaffold for optimal photoswitching regarding the excitation wavelengths were optimized.^[177] This is the (only) discrete optimization so far with OGOLEM, meaning the sampling (besides local structural optimization) within a pre-elaborated database of functional groups on predetermined scaffold places.¹ Multiple (well-known) benchmark functions² in the regime of global optimization were optimized, in order to show that actually most of these are trivially solvable with OGOLEM.^[394] The latter task can also be assigned to the parameter optimization regime. However, not *parameters* of any other function are to be optimized with respect to a target, *e.g.*, the minimal deviation from a reference, but the benchmark function is the objective function itself. Global optimization of small water clusters ($n = 3$) with sodium marked the first optimization on a higher level of theory (DFT) with experimental comparisons.^[395]³ OGOLEM was used for the fitting of Gaussian model potential's parameters for allosteric anion binding in cage complexes,^[398] which was extended to more complex cages thereafter, also including comparisons with experiments.^[399] An explicit solvation shell of H₂O mixed with additional implicit outer-shell water was globally optimized around separate frames of a Menshutkin reaction.^[400] Also larger water clusters ($n \approx 30$) with generated infrared spectra were investigated in line with corresponding experiments.^[401]

Moreover, regarding method development research, “graph-based directed mutations” were focused first of all.^[402] These are certain problem-specific operators that count promising stabilizing interactions/connections and introduce meaningful jumps through the search space for the CSO problem. Both in line with method development and with the aforementioned water cluster experiments, more complex objective functions than the usually used simple energy expression for CSO was illustrated in Ref. [403]. Furthermore, a new (multiple) abstraction layer for arbitrarily mixed distributed and shared memory computations was published in Ref. [404] using an asynchronous server–client distribution protocol of GA jobs with intrinsic error-safety against hardware failures and transmission problems.⁴ This **remote method invocation (RMI)**-based parallelization layer and its robustness was shown in Ref. [405]. Here, the smaller and bigger scheduling gaps that occur on **high-performance computing (HPC)** clusters were automatically filled up with OGOLEM jobs that are also killed again if the resources are needed in other jobs without deteriorating

¹ Note that this was intentionally a proof-of-concept study. Hence, the database was merely a small list of meaningful functional groups in this case.

² The benchmark functions such as Schwefel's, Ackley's, Schaffer's functions, etc., are given in the same Ref. [394] or also in Ref. [29].

³ Note that all optimizations in OGOLEM's workflow are completely independent from the level of theory in general. Many empirical potentials are implemented in OGOLEM, but for all other needed levels of theory a respective external package is called. From this perspective, it might seem peculiar that other programs explicitly emphasize their used level of theory.^[396,397]

⁴ This additionally enabled new features at the same time such as a multiple “island” GA optimization—each worker with a different GA setting competing within the same overall population.

the OGOLEM optimization. Further hybridization with other ingredients were developed, *i.e.*, mutation operators working on a specific spatially local Cartesian neighborhood in a picture of a “heat pulse” impacting a cluster, which can be seen as phenotype mutation perturbation and relaxation.^[406]⁵ Global parameter optimization of local pseudopotentials as they are used in orbital-free DFT was approached which accordingly also led to more complex objective functions with multiple different properties of the solid and/or liquid phase during the fitting procedure.^[407] Recently, structure optimization of rare-gas LJ clusters influenced by heterogeneous (fixed) surfaces also consisting of rare-gas atoms was illustrated.^[408] Similar surface-attached clusters, this time on gold, were investigated to explore the self-organization of molecular tether molecules.^[409]

Somewhat currently (still) detached from the main OGOLEM line (see the next Section 3.2), the current author of this Thesis started with the development of interfacing REAXFF as implemented in sPUREMD^[410,411] and OGOLEM in his Master’s Thesis, for full-fledged global parameter optimization of the former reactive force field, including also many different properties in the objective function.^[240] As some nuisances of the linear parallel scalability of that implementation have been noticed then, this was solved by very low-level memory-scheduling and some OGOLEM designs and published afterwards already as part of this Thesis. This is discussed in Chapter 4.^[244] The same implementation was used for fitting REAXFF potentials to high-level multi-reference data for studying disulfide mechanochemistry,^[138,247] *i.e.*, the breakage of disulfide groups by an attached mechanical force.⁶ Focusing on different algorithms for secondary order parameters, *i.e.*, niching (*vide infra*), we have shown that the latter are indeed needed for the hard LJ cases for conserving structural diversity during the optimization. Without much fine-tuning, meaning if used at all, such hard cases are certainly solvable. Besides, the algorithm for niching, which was already used in Refs. [240, 244, 247] and afterwards in this Thesis, is described briefly below (see Section 3.5.3), whereas the actual niching details in the LJ CSO case are discussed in Chapter 5. Last but not least, the GOCAT concept was introduced by showing many qualitatively different settings of partial charges around a Menshutkin reaction for optimal reaction barrier lowering. This follows in Chapter 6. Furthermore, unpublished extensions of this framework to adaptive re-optimizations of the MEP during the GOCAT optimizations are given in Chapter 7.

3.2 Overview

As the current author already started in 2014 during the Master’s Thesis^[240] to develop all kinds of new features for OGOLEM, the code base has increased considerably since then, see Table 3.1 on the following page (all of the source lines of code (SLOC) were computed by cLOC^[412] in this Table). At the time of writing, this still is a *local* branch of the published

⁵ As illustrated in Section 2.1.2.2, any unary operator working on a candidate solution and returning a new candidate solution can be termed mutation in the terminology of a GA: Be it a little MC step or a full-blown multi-step protocol.

⁶ This also included a well-known “regularization” protocol, the *early stopping*^[353,354] with training/test set splits, for avoiding overfitting—though, such schemes are not implemented in OGOLEM yet.

work of DIETERICH in Ref. [413], but actually merged with the default branch until October 2016 and therefore including the updates of Ref. [404].⁷ Also, notable work was put into generifying common classes for both **CSO** as well as **GOCAT** optimizations, which is mentioned here to point to the fact that some parts of the new **SLOC** found in Table 3.1 are not *purely* due to new features but to changed/adapted versions of the default **OGOLEM**. Further work on merging this current branch with the default one is scheduled into the near future. Also note that the pure number of **SLOC** is by no means descriptive of the actual content implemented and is neither comparable between different programming languages nor between various “dialects” within the same language. Therefore, this solely shall illustrate the extent of the changes or adaptations made so far. One implementation example in the realm of software engineering is discussed in Section 4.3.

Now, we will start a more detailed discussion of **OGOLEM** and the new features.

Table 3.1: Size of the code base developed for this Thesis.

Branch	Language	Files	SLOC
OGOLEM (default) ^a	Java	497	70000
	Java	910	150000
OGOLEM (Thesis) ^b	C ^c	31	12000
	Python ^d	84	16000

^a This is the code downloaded from Ref. [413] on 03/07/2019. Note that external libraries that are compiled to byte-code are not included here, which encloses also some content of DIETERICH, e.g., **EVb-QMDFf**,^[227,264] a scala implementation of **TTM3F**^[414] and some others, besides the usual logging, unit test, linear algebra packages, etc.^[413]

^b This is the local branch with *all* extensions that started during the Master’s Thesis of the current author and that progresses to the work in this Thesis, including Chapters 4 to 7.

^c This is **REAXFF** as implemented in **sPUREMD**^[410,411] and adapted for Ref. [244], cf. Chapter 4. Note that this not *strictly* belongs to the same code base, but it is distributed *separately* due to the **GNU GPL** vs. the four-clause **BSD** license (without copyleft) of **OGOLEM**.

^d This is also external to **OGOLEM** and is the basis of all clustering/analysis for **GOCAT** optimization as done in this Thesis. Here, all types of programs with heavy use of all the powerful libraries for scientific computing available in Python were written, utilizing the common Python ecosystem: **NUMPY**,^[415] **SCIPY**,^[416] **MATPLOTLIB**,^[417] **PANDAS**,^[418] **SEABORN**,^[419] **STATSMODELS**,^[420] **SCIKIT-LEARN**.^[367] All (non-molecular) Figures in this Thesis were created with **MATPLOTLIB**, except for Fig. 3.8, which was generated with **GNU PLOT**^[421].

3.3 OGOLEM’s Genetic Algorithm

The main **GA** cycle of **OGOLEM** is pictured in Fig. 3.1 on the next page (compare with Algorithm 2.3 on p. 29). In 2006, **BANDOW** and **HARTKE** already established the so-called “pool-**GA**” instead of the more commonly used paradigm of generation-based **GA**.^[422] This constitutes also the core-part of **OGOLEM** as now almost all parts of the cycles are trivially independent of each other. By comparison, the usual bottleneck of a generation-

⁷ However, in, e.g., Ref. [407] substantial changes to the adaptive part (responsible for parameter optimization) took place that are not orthogonal to the ones of Ref. [244]. So, without the possibility of novel features for this Thesis, the hence necessary work of *manually* merging and re-implementing some bits was postponed to a near future.

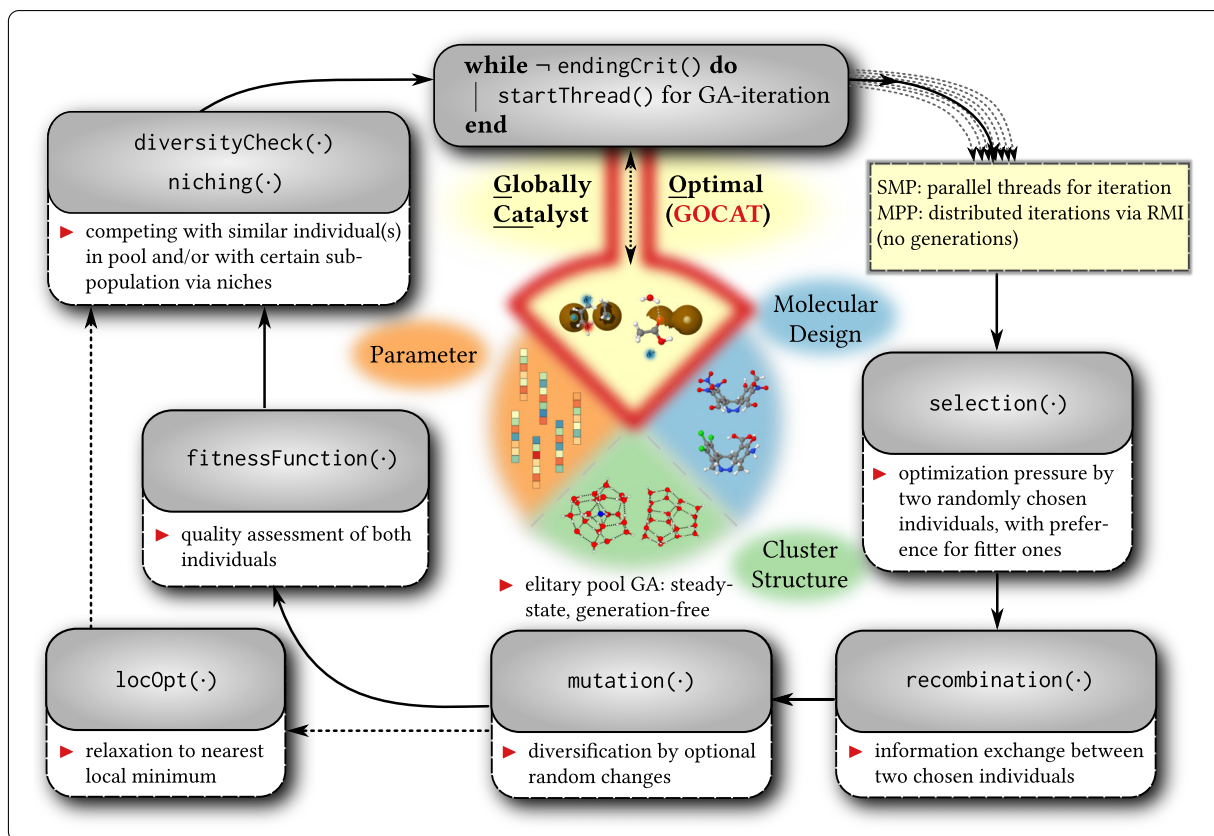


Fig. 3.1: After build up of the pool as initialization (not shown), the main OGOLEM GA optimization cycles are illustrated here for the four illustrated optimization tasks discussed in Section 3.1. Some further acronyms appearing here are: **symmetric multiprocessing (SMP)** and **massively parallel processing (MPP)**, which are discussed again in Section 4.3.

based algorithm is imposed by the discrete exchanges and waiting times for swapping (or merging) complete populations all at once. In contrast, in the pool version, the serial bottleneck reduces to drawing two individuals, *i.e.*, the selection, and merging one new individual with the pool such that the updates happen in a piece-wise fashion. As this latter computation time is usually negligible compared to all the other steps (crossover, mutation, fitness assignment), we usually reach linear scalability for *strong scaling* of each optimization task.^{[389]⁸} That is, a linear decrease of the wall clock time is observed which is proportional to the increase of the number of cores involved in the computations. This topic will be readdressed in Section 4.3.

In OGOLEM, we always use a fitness-sorted pool where each individual gets a rank based on its index within the pool. Then, in the most common setting used also in this Thesis for `selection(.)`, a non-parametric selection happens *via* this rank order. Here, a Gaussian-distributed random number around the best rank, r_0 , with one side of the distribution to the worse rest of the population is drawn. This delivers one good candidate solution; the second is often just uniformly selected at random. These individuals are again tweaked

⁸ In the literature, this is simply called *steady-state GA*,^[29,156] and also the structure of using a “pool” without generations for the parallelization can be traced back to Ref. [423] as described in our Ref. [422]. Later, this idea of an generation-free GA, especially as an advantage for the parallelization, have spread to other packages, *e.g.*, Refs. [424–428].

by binary and unary operators to produce two new children, *via* `recombination(·)` and `mutation(·)`. The single quality assessment with `fitnessFunction(·)` can alternatively also be extended to a full-blown local optimization *via* `locOpt(·)` instead. Then fully relaxed individuals within their basin of attraction are returned with respect to the fitness. In many settings, analytical gradients are not readily available since the final fitness value is a result of a more complex simulation, which is the case for REAXFF parameter optimizations, or a composed measure of many ingredients and even heuristics, which is the case for GOCAT design; this topic is readdressed in Section 3.6. Then, for this purpose, explicit gradient-free optimization instead of labour-intensive numerical gradient approximations can be used, leveraged by algorithms such as BOBYQA^[144] (with restraints) or NEWUOA (without restraints).^[145] If gradients are available as in CSO, the quasi-Newton LBFGS is the best bet. Next, the better one of the two children is compared to the pool and the decision is made whether to replace one other individual. This is only done if the current fitness is better such that our algorithm is strictly *elitary*. This means that a deterioration of the solutions cannot occur in any single iteration, which is a more exploitative variant of a GA. This guarantees that the population will converge. If then the global minimum is not (yet) found (which is never guaranteed), this is dubbed *premature convergence*. Thus, further secondary quality measures of enforcing some diversity within the pool are of utmost importance. This happens in the final step summarized in `diversityCheck(·)` and `niching(·)`.⁹ In general, we check multiple criteria such as having a different fitness from the rest of the pool and other structural intrinsics.¹⁰ In a chemical context, HARTKE introduced the concept of *niching*.^[167] In a neutral, less bio-inspired way this can simply be called *order parameter* which usually encodes further intrinsic information of a candidate solution for the problem, for instance, the configurational setting of a GOCAT. The *static* niching version divides the overall search space into confined areas where each candidate solution competes with just the subset within the same area. Alternatively, a *relational* version of niching was introduced by the current author^[240] which dynamically changes with respect to the overall current pool and catches minimal metric distances between clustered subsets of the pool. This whole topic is the main issue of the publication presented in Chapter 5. In the end, because of the diversity enforcement, the resulting pool will consist of qualitatively different candidate solutions, where each will represent usually a local optimum.

3.4 Further New Capabilities

In the following, a rough enumeration of the basic extended capabilities of the OGOLEM package developed during this Thesis will be given. These points only consider the most

⁹ Similar schemes such as *sharing* or *clearing* in the context of diversity conservation can be found in Ref. [29].

¹⁰ Naturally, the same fitness can also occur at different coordinates on the fitness surface when the minima are “degenerate”. In CSO, with an increasing number of particles there is an increasing number of local optima having the same energy in different configurations, by chance. In other problems such as the GOCAT design the heuristics and accumulation of the fitness makes the final fitness value probably ill-posed or underdetermined; see the discussions around the RESP charges in Section 2.1.1.2 and in Section 6.3.2.

dominant features regarding **GOCAT** optimization; some other diverse improvements are not explicitly stated here.

- Packing operator on different spaces (sphere, ellipsoid, etc.), including a **vdW** surface: Usually a numerically tessellated exposed surface of all atoms of all frames is created such that the initial packing can take place on this surface directly. By this, the probability of selecting specific atoms is also weighted by the *exposed* surface area.
- Restraints handler in order to map to certain surfaces again after each operator application: After/before each application of any operator, all charge entities can be moved or changed in order to fulfill all defined model constraints again; this also includes charge value constraints. Either this can be used as a restraint during any other type of optimization as, *e.g.*, the gradient-free **BOBYQA** local optimization of the full fitness function, or this is used as a stand-alone “**FF**” and part of a *chain* of operators to satisfy the constraints. In this way, an arbitrary operator can work on a **GOCAT** which is corrected afterwards to fulfill its model constraints again.
- About ten different niching algorithms, including a plethora of sub-algorithms—also heavily re-using some similar implementations present already in the **CSO** regime and parameter optimizations (such as overlap nicher, substructure nicher, *abstract* vector-based nicher, etc.). For **GOCAT** design, there are different protocols, for instance, Voronoi tessellations in the Cartesian domain, **ESP**-based similarity measures (discrete or continuous), **CM**-based measures and others. One **ESP**-based discrete relational niching will be discussed below in Section 3.5.3 and the **CM**-based niching is used and explained in Chapter 5.
- Six different mutation operators—again often based on generified versions of pre-existing **CSO**-operators—, *e.g.*, standard **MC**-based ones in the Cartesian domain, an exchange mutation operator, **MC** steps for the charge values, etc.
- (Just) one “mighty” phenotype crossover operator, *sweden* (explained in Section 3.5.1),¹¹ in addition to the already available other ones that did not have to be changed. Another operator, *canada* (explained in Section 3.5.1), is rather a mutation operator at the moment but could easily be generalized to wrap anything, also other arbitrary operators.
- Several *utility*¹² routines were implemented as for instance: Translation between levels of theory using the “distances” or “**vdW**” protocol (described in Section 6.3.2); a separate **NEB** reaction path utility implementing linear/non-linear interpolations, improved/eb/dneb tangents, the adaptive **NEB** version, **CI**, etc. (all that was mentioned in Section 2.5.1).

¹¹ Because of the resemblance to the **CSO** and its sense there, the operator was named the same.

¹² General utilities define an own execution entry without starting the main **GA**.

- Electrostatics utility routines for scalar φ_{ESP} potentials and OCFs.^[79]
- Utilities for re-evaluating the GOCATs, also on different levels of theory: Re-optimization of stationary points, using either OGOLEM's internal algorithms or external ones, for the stationary points (R, TS, P) and frequency checks (for which external programs are used).
- Five backends for GOCAT specifically implemented: MNDO,^[429] MOPAC,^[226] ORCA,^[430] XTb2,^[228,229] EVB-QMDF.^[264]¹³
- Different additional local optimization algorithms such as FIRE and LBFGS without line searches and both with different robustness techniques such as: Restarts at fluctuating gradient directions, rescaling of step-lengths (if they become too large), history of all former steps if the optimizations become unstable and restarting/returning proper ones. This is all needed for non-conservative NEB-forces, essentially, and for optimizing structures in (extreme) electric fields.

3.5 More Detailed Operator Descriptions

In this Section, some of the used and implemented GA operators are described in some more details. These descriptions are limited to *exactly* the operators that are used in later Sections of this Thesis. Hence, in Section 6.3.1 all these backgrounds are needed when these operators are benchmarked. In OGOLEM, there are some operators available that work on a somewhat more abstract genotype, although no real genotype-to-phenotype mapping, which is used in the traditional first GA implementations or other application contexts, formally takes place. These more abstract operators have the clear advantage to be applicable in all kinds of optimization problems so that the same generic algorithm can be used for, e.g. CSO, parameter optimization or GOCAT design. However, knowing the problem setting at hand, more concrete and problem-aware algorithms can be implemented tailored to specific applications that can bring in additional efficiency, while at the same time giving up the generality or even bringing in some bias.¹⁴

3.5.1 Recombination

One recombination operator which rather belongs to the genotype operators is illustrated in Fig. 3.2 and is called portugal.¹⁵ As being of this generic type, it belongs already to the established assortment of operators and was not newly implemented in this Thesis. portugal works on any array or string of arbitrary type that can be cut at different points and spliced together. Thus, one optimization task must polymorphically decide what a

¹³ An internal OGOLEM backend for EVB-QMDF actually was implemented by DIETERICH in the current version. Older versions fell back to original FORTRAN code the current author adapted, which is not mentioned in Table 3.1.

¹⁴ In this context, the well-known *no free lunch* (NFL)^[431] theorem is to be re-stated, which will be delayed until Section 6.3.1 where these operators are benchmarked.

¹⁵ In OGOLEM, most operators traditionally carry country names.

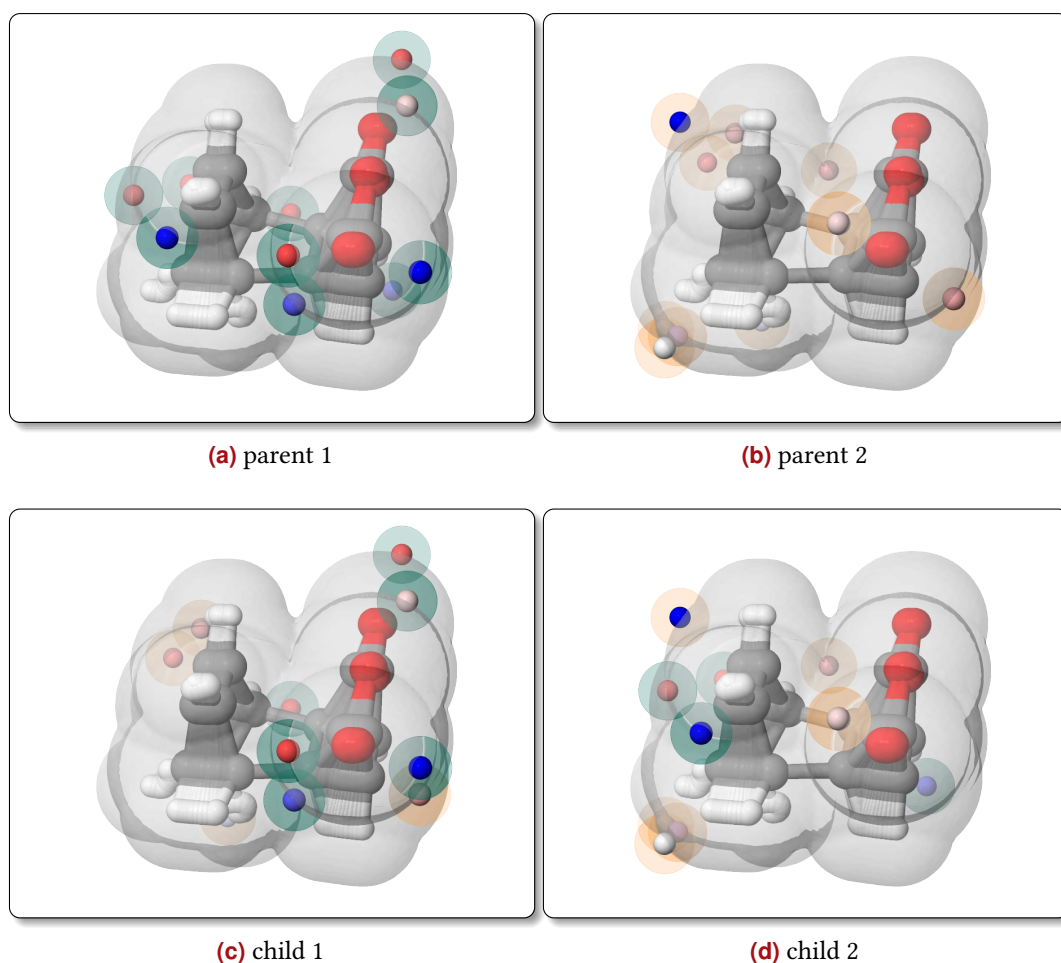


Fig. 3.2: Typical genotype recombination operator, portugal: $N_{\text{Ch}} = 10$ charges between both parents are redistributed by swapping subsets of the charge entities after cutting the whole list at some points (3 in this example). The resulting children's charges are perfectly placed on the **vdW** surface, but the resulting child is not yet neutral; the **vdW** surface is sketched as white translucent common manifold of 18 frames of the **MEP** of a **Diels-Alder** reaction between maleic anhydride and cyclopentadiene. (This reaction will be the central topic of Chapter 7.)

meaningful entry of such an array is. Certainly, this is not just *one* of, *e.g.*, the four floating numbers corresponding to the position vector and partial charge value but the combination of these, *i.e.*, $\{\mathbf{r}_i, q_i\}$. In other words, each meaningful subset of intrinsic coordinates of one charge entity is treated as a whole and exchanged in this way.¹⁶

In Fig. 3.2 one arbitrary recombination step of this type is shown, where the partial charges are highlighted with halos in order to see where the charges ended in the children after redistribution. Such an operator that clearly swaps parts of the array of partial charges is known to be one which is quite exploitative, jumping from corner to corner of a hypercube in the search space, sampling the *combination* of what already is present in the parents without creating new numbers at all.¹⁷

¹⁶Hence, actually by already defining a representation what/where a meaningful cutting point in an array of coordinates is brings in some phenotype character, actually. This elderly dichotomy of genotype vs. phenotype is therefore somewhat fuzzy nowadays in these chemical optimization problems.

¹⁷Assume that we have 3D vectors as points in search space. Then two vectors to be crossed would each form

On the other hand, an operator rather belonging to the phenotype ones is illustrated in Fig. 3.3 on the following page and is named *sweden*. This one is related to the first 3D cut-and-splice operators from Refs. [167, 180, 181] and yet it was adapted for **GOCAT** optimization during this Thesis. In this case, a 2D plane is generated with a random orientation and support vector. Then the parts, which do not necessarily have to be equal halves, are redistributed to generate the children. Thus, *sweden* lets full parts of the candidate solutions in the Cartesian domain intact. In **CSO**, this translates to exchanging possible *stable* cluster parts with most of their meaningful interactions to neighbors except for the distortions due to the cutting plane. In contrast, the *portugal* operator for **CSO** would just swap atoms or molecules themselves. In **GOCAT** optimization, there is no such thing as a “stable” charge half with meaningful interactions between the charges themselves. The purpose rather is to exchange parts of, e.g., a catalytic fit **ESP** at the bottom of one parent with the (hopefully acting synergistically) other top part of another parent. This operator then orients along the 3D space and does not only cut 1D vectors of charge entities as in *portugal*. Actually, there is a chance that both operators would result in the same child if they were applied to the same **GOCAT** in comparison. Therefore, we deliberately introduced even more search space coverage by this operator with using *different* orientations for the cut planes in both parents. In this way, a bottom part can be exchanged also with another bottom (or whatever) part of another **GOCAT** that will not perfectly fit onto the exposed **vdW** surface unless being completely spherical or by mere chance. Moreover, for finding each orientation some Cartesian translations due to center of mass transformations to the origin of the coordinate system of all charge clouds take place. For clusters in the **CSO**, in contrast, one could try to find a type of meaningful distance between both halves before combining them (when the overall shapes are not complementary). For the **GOCATs**, however, this is not meaningful in the crossover step yet and is delegated to the later restraint handler. Consequently, this leads to a subtle additional “noise” effect and results in some further explorative nature of the algorithm. In the **CSO** regime, one can even make a subsequent local optimization, just to find out at which distance and orientation the two parts should be placed together, *i.e.*, one can find the optimal *relative* coordinates of the parts. For **GOCATs**, just the emergent **ESP** matters and the charges after splicing together the halves will probably already be slightly off the **vdW** surface. Hence, some additional noise by not defining “standard” relative placements can be accepted, too.

This is also shown in Fig. 3.3 on the next page. For better grasping the cut-plane methodology, bigger **GOCATs** with more charges ($N_{\text{Ch}} = 50$) are shown and the resulting children both have charges that are not yet perfectly placed onto the **vdW** surface. This remapping as well as (if needed) overall charge value constraints, $q_{\text{sum}} = \sum_i q_i$, are always executed after each step, usually; this can be done by the restraint handler mentioned above in Section 3.4.

a corner of a 3D box and each child would be located at another corner of that same box. Yet, no *volume* inside that box would be sampled (e.g., compare with Ref. [29, Fig. 29.6, p. 338]).^[29,156,240] In Ref. [240], the current author introduced the average crossover operator for real vector spaces for parameter optimizations, *arctic*, that also generates new averaged numbers for the vector elements and does not only re-combine the already available ones.

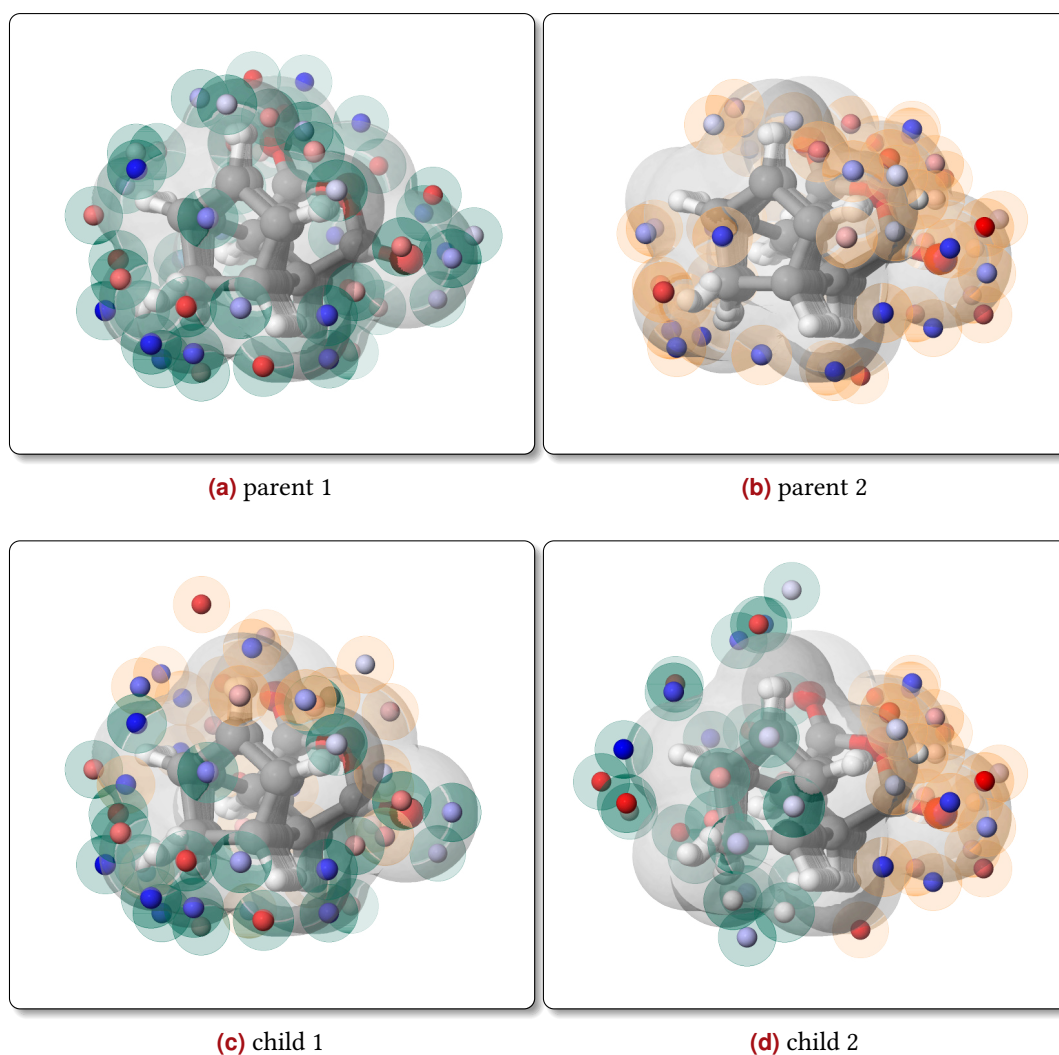


Fig. 3.3: Typical more phenotypical operator: $N_{Ch} = 50$ charges between both parents get redistributed by creating two different 2D cut planes and exchanging the parts. The resulting children’s charges are *not* perfectly on the **vdW** surface as the surface is not spherical; neutrality and re-mapping onto **vdW** happens thereafter.

In conclusion, sweden firstly cuts directly in 3D spaces by using arbitrarily set and oriented planes, but secondly, by possibly splicing together arbitrary parts, the **vdW** surface can usually not be matched, which brings in the additional “fuzziness” or explorative nature intended. Hence, even if two selected parents were the exact same—the parents are usually selected with replacement—,we could generate different children.

3.5.2 Mutation Operator: canada

The operator named canada is meant to introduce some more problem-specific behavior into the **GOCAT** optimization. Therefore, one could term it “phenotype mutation” operator, similar to, *e.g.*, the purpose of the cut-and-splice binary operator in 3D Cartesian space, sweden, described above, or similar to the phenotype “heat-pulse” operator^[406] mentioned in Section 3.1. In canada, one individual is mutated to result in another **GOCAT** of different

coordinates, $\{ \mathbf{r}_i, q_i \}$, and yet to show a similar overall φ_{ESP} at specific frames; these frames usually belong to a subset of the discretized reaction path, including the most important three stationary points (R, TS and P). This is illustrated in Fig. 3.4 on the following page (all data here is exemplified using the Menshutkin reaction on PM7, as studied in Chapter 6).

First, the starting individual shown in Fig. 3.4(a) is mutated by applying a *chain* of different other mutation operators subsequently. In Fig. 3.4(b), this was done 200 times and all the resulting GOCATs are superposed in that Figure. In detail, each charge position, \mathbf{r}_i , is moved in a random direction with a uniformly sampled displacement vector of $\mathbf{r}_i^{\text{shift}}$, in this example, of 1 Bohr (maximum) length. Simultaneously, the charge values (actually, here a subset of these), q_i , are also shifted under a Gaussian distribution centered at the current value before, q_i^{ref} , with standard deviation of $\sigma = 0.1 e$. Hence, each partial charge will be kicked off into the Cartesian vicinity and might carry a slightly different charge. Note that this intermediate state is illustrated in Fig. 3.4(b), termed “displaced” in the caption, as the vdW surface now is left transiently. Next, a gradient-based local optimization follows, by commonly using the LBFGS algorithm, with respect to the ESP-difference as objective function, *i.e.*, Eq. (2.14) on p. 18. This is carried out under all the constraints described in Section 2.1.1.2 on p. 18 and in consideration of the proper symmetry, here of C_{3v} . In other words, this kicks the coordinates of a GOCAT over some nearby barriers and relaxes them to the new nearest local optimum with regard to the summed error of ESP-differences before and after the mutation step.

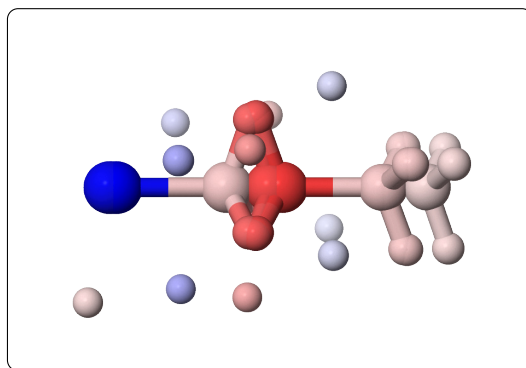
The result of this optimization and hence of canada’s complete mutation step can be seen in Fig. 3.4(c); this is illustrated for the 200 separate operator applications, and all charges are now relaxed to the nearest local optimum of the ESP objective function. Furthermore, all charges reached charge neutrality of $\sum_i q_i = 0$, the vdW surface radii and minimal distances of 1 Å between mutual charges again. Usually, already such small displacements—in fact, the concrete domains of each charge have not changed very much from Figs. 3.4(a)–(c)—will not result in $\|\Delta\varphi_{\text{ESP}}\|_2 = 0$ (*cf.* Eq. (2.14)).

Figs. 3.4(d)–(e) show again 200 separate applications of canada, but this time with bigger mutations of the chain of Cartesian and charge displacements of $\|\mathbf{r}_i^{\text{shift}}\|_2 \leq 3$ Bohr and $\sigma = 0.2 e$. In this example, quite different embeddings can be reached in Fig. 3.4(e), as can be seen by the almost continuous surrounding.

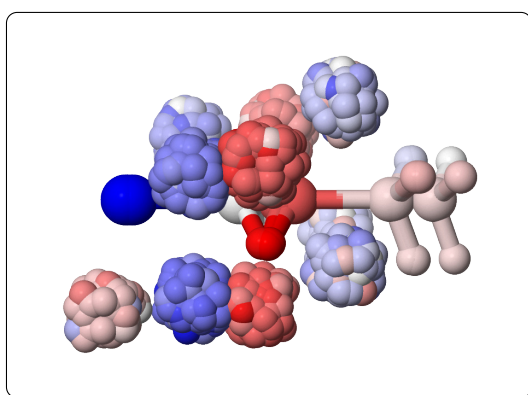
canada was intended to be an “end-game” operator, using an iteration-based protocol by mixing in this operator after some initial GA progression. That is, after some iterations, already catalytically active GOCATs will be found and constitute the population. Then, one such GOCAT can be tweaked by canada to sample a similar one with respect to the ESP. However, as we have observed (see later Section 6.3.1), some canada in the mixture¹⁸ of different mutation operators is already beneficial right from the beginning of the GA.

Corresponding energies and gradient norms are given in Fig. 3.5 on p. 83. Energy profiles for the small canada mutations of Fig. 3.4(c) are pictured in Fig. 3.5(a) and paths *without* any ESP re-optimization in Fig. 3.5(b). canada introduces the ESP-aware steps, whereas

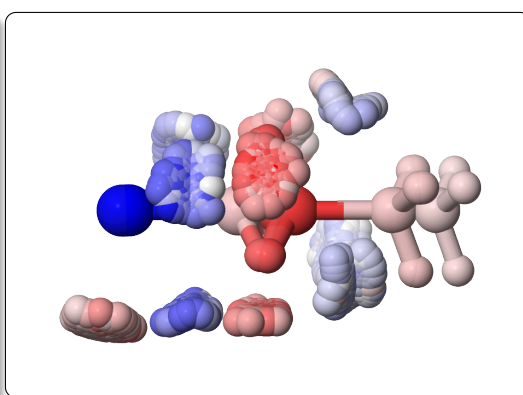
¹⁸Usually, exactly the two mentioned mutation operators in the *chain* (Cartesian or charge displacement) wrapped by canada are separately in the mixture, too, without any ESP optimization.



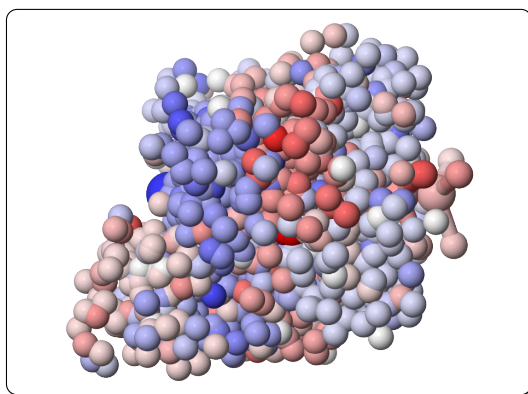
(a) starting individual (r_0)



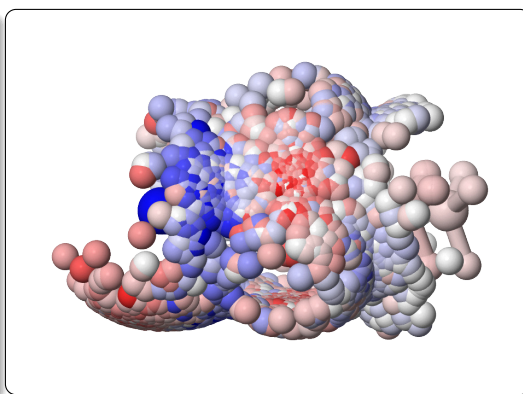
(b) small canada mutation: displaced



(c) small canada mutation: optimized



(d) bigger canada mutation: displaced



(e) bigger canada mutation: optimized

Fig. 3.4: Two cases for the canada mutation are illustrated: In Fig. (a) the starting GOCAT is shown with $N_{\text{Ch}} = 10$ where the charges that carry a charge value q_i and the atoms are colored from red to blue in $q_i \in [-1, +1] e$ and $\varphi_{\text{ESP}} \in [-27.0, 27.0] \text{ kcal mol}^{-1} e^{-1}$ at the $27 = 3 \cdot 9$ atoms of R, TS, P, exemplified for the Menshutkin reaction (*cf.* Chapter 6): $\text{Cl}-\text{CH}_3 + \text{NH}_3 \longrightarrow \text{Cl}^- \cdots \text{H}_3\text{C}-\text{NH}_3^+$. In Figs. (b)–(e), 200 different separate mutation applications are superposed, corresponding to 2000 separate point charges in total.

the latter energies of Fig. 3.5(b) are based on the exact same chain of Cartesian and charge mutations, but without optimizing and ESP afterwards; in this case, just the plain constraints, such as overall neutrality and the vdW surface attachment, are optimized as usual after every operator application. Apparently, canada leads to mutated GOCATs having energetic profiles that are very similar to the starting individual r_0 as almost all lines lie on top of each other. Without the ESP optimization, but still using the same small chain of mutation steps, the profiles are very erratic, as shown in Fig. 3.5(b). In the former case with ESP optimization, Fig. 3.5(a), about 58/200 GOCATs have a fitness that is smaller than the starting one without any GOCAT. Since all energies are clearly very similar, the fitness directly scales with the gradient norms; the composition of the fitness function considering energies and their gradients is explained in Section 3.6. If the gradient norms become too large at the indicated three frames ($\gg 10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$), the GOCATs will not successfully compete with the population. In the latter case of no ESP optimization, Fig. 3.5(b), there are some individuals without any barrier decrease at all or some with large gradient norms at the former stationary points. These worse individuals might connect any non-stationary point on the PES such that the energy barrier decrease is an *illusion* without having an approximate MEP. Here, 20/200 GOCATs are in a catalytic region regarding the fitness.¹⁹

The energies follow strictly the applied (and very similar) ESP influences at the atoms after canada, whereas the gradients are not that uniform, generally. Note that just three frames and only the core atomic positions are included in the scalar ESP optimization. There is no quasi-continuous grid of such ESP values included in a more extended Cartesian space around the reaction. Especially due to the classical QM/MM coupling applied here (cf. Section 2.4.3), the energies correlate exactly with the subset of ESP values at the core atoms, while the gradients are dependent on the precise Cartesian positions of the charges. In the very symmetrical r_0 , these gradient components stemming from different opposite charges compensate each other, but in many other canada mutated individuals, this is not the case.

However, this general performance is not too problematic at all because this sub-space ESP plays a role of another heuristic within an operator and apparently brings in some performance enhancement. This will be readdressed during some benchmarks in Section 6.3.1.

Finally, the separate ESP values are given as heatmaps in Fig. 3.6 on p. 84. In Fig. 3.6(a), the φ_{ESP} is almost the same at each of the atoms included as reference for the ESP optimization, as intended, but with the aforementioned very small differences of $\|\Delta\varphi_{\text{ESP}}\|_2 \neq 0$. Without such re-optimization, the ESP values are arbitrary after the same small mutation chain and given in Fig. 3.6(b) (note the bigger scale for φ_{ESP}).²⁰ The best rank, r_0 , is also plotted in each case at the top of the ordinate axis and sorted with respect to the fitness downwards. Thus, the number of GOCATs that lie in the catalytic region for Fig. 3.6(b) are oriented

¹⁹The attentive reader might have noticed the shift of the minimum of the gradient norms of some paths and of r_0 in particular. This is the expected trend of already setting in non-vertical impacts on the path that is possible in that simple Menshutkin reaction. This trend is discussed thoroughly in Chapter 6.

²⁰The mean standard deviation of ESP of Fig. 3.6(a) is $0.24 \text{ kcal mol}^{-1} \text{ e}^{-1}$ and the one for Fig. 3.6(b) is $35 \text{ kcal mol}^{-1} \text{ e}^{-1}$ with some outlier values also lying in $[-150, +120] \text{ kcal mol}^{-1} \text{ e}^{-1}$ that are clipped in the heatmap.

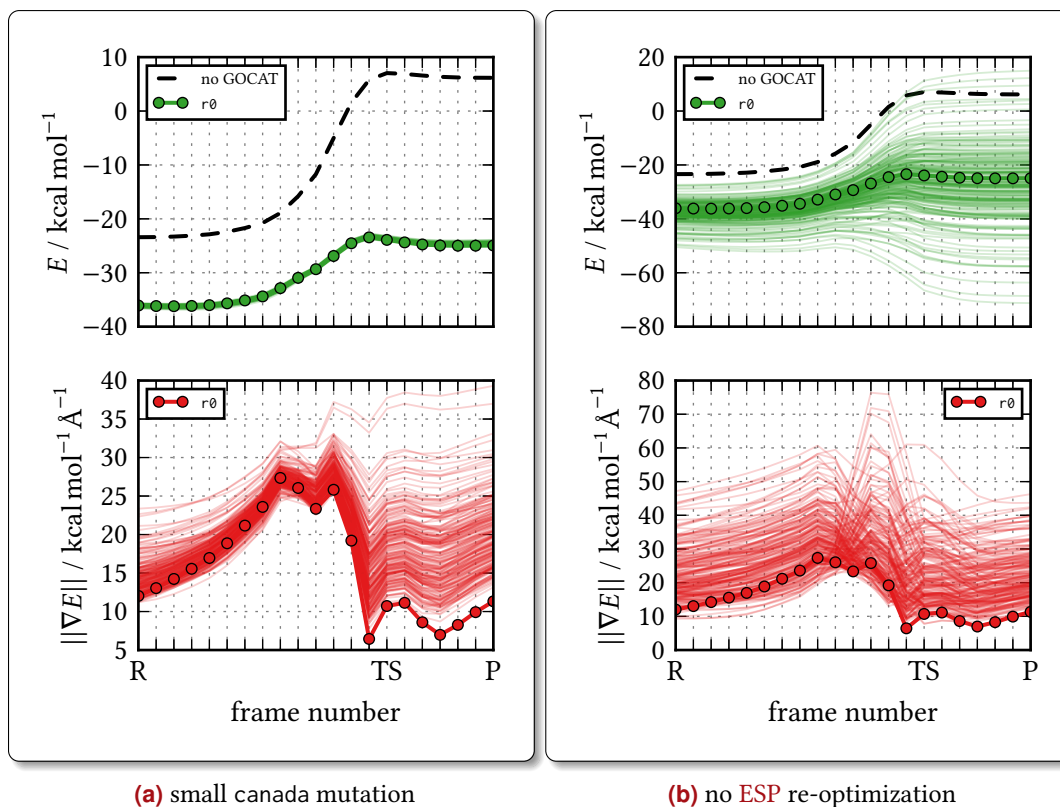


Fig. 3.5: Superposed energy profiles and gradient norms for *canada*. In Fig. (a), the energies of the separate *GOCAT*s of Fig. 3.4(c) after the small *canada* steps are shown. In contrast, energy profiles of *GOCAT*s without *ESP* optimization, *i.e.*, only after the plain *chain* of the other mutation operators but without *canada*, are illustrated in Fig. (b); The black line is the profile without any *GOCAT*, whereas $r\theta$ is the best *GOCAT* with $N_{Ch} = 10$ ever found for this reaction (*cf.* Fig. 3.4(a)).

at the top of that Figure, while the worse individuals with apparently highly varying *ESP* follow downwards.

The trend now using even bigger mutations as in Figs. 3.4(d) and 3.4(e) is as expected (but no further Figures are shown here): Less individuals will be in a catalytic region since quite huge jumps are introduced. Again, the final *ESP* reached after *canada* is rather similar to the starting individual, as intended, but with bigger deviations than in the smaller mutation case discussed above. In practice, both *canada* versions (smaller and bigger steps) are used in the *GA* operator protocol.

3.5.3 Niching

The introduction of an additional order parameter into the *GA*, *i.e.*, niching, was already mentioned in Section 3.1. For the *GOCAT* design, only one but the most often used protocol is illustrated in Fig. 3.7 on the next page. In this case, the *ESP*, φ_{ESP} , is calculated at the core atoms for all individuals in the population. In each call of *niching*(\cdot) (see Fig. 3.1 on p. 73) the new child individual is then compared to *each* individual of the current pool. One such comparison is shown in Fig. 3.7. These are binary comparisons here that check whether two φ_J^{ESP} values of two different individuals are formally equal or unequal at the

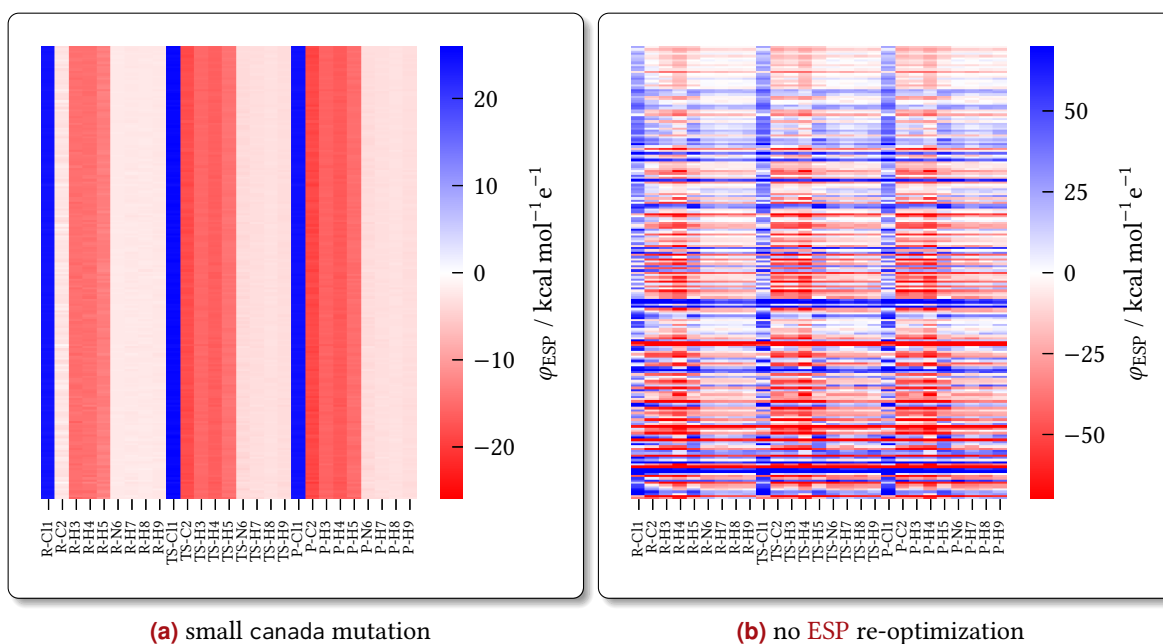


Fig. 3.6: Heatmaps of the ESP at the 27 usually included atoms of R, TS and P frames for the Menshutkin reaction: The case in Fig. (a) with ESP optimization corresponds to Figs. 3.4(c) and 3.5(a). The case in Fig. (b) without any ESP optimization (plain mutations) corresponds to Fig. 3.5(b). The 200 GOCATs are plotted (ordinate) against each $9 \cdot 3$ atoms for the three frames (abscissa); these are denoted here with, e.g., “R-Cl1” standing for “chloride atom 1” of the R frame, etc. (Note that for the current context neither the exact atom enumeration nor the ESP values are important; the overall similarity of each ESP that is reached is the focus here.)

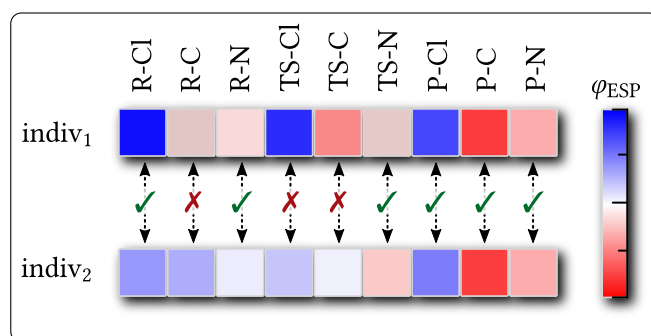


Fig. 3.7: Simplified illustration of a discrete φ_{ESP} -based niching. Two ESP potentials at selected core atoms are compared atom-wise and counted as being “unequal” (X) or “equal” (✓). The atoms are denoted with, e.g., “R-Cl” for the chloride atom at the R frame (similar to Fig. 3.6). This is one protocol, besides many others, for the same ESP vectors.

atom J . The allowed difference, $\Delta\phi_{\text{thresh}}^{\text{ESP}}$, is set beforehand as a simple threshold parameter. If the difference at one atom is greater than, $\phi_J^{\text{ESP}} > \Delta\phi_{\text{thresh}}^{\text{ESP}}$, this is considered as one inequality. By counting the number of equalities and inequalities, the individuals are then assessed to be similar when too many equalities occurred. In this case, the two individuals are considered to populate the same niche. Otherwise, they are treated as being dissimilar, *i.e.*, they populate two different niches. The number of maximal (in-)equalities, $N_{\text{deviations}}$, is also pre-set. Note that besides these integer-based atom-wise binary comparisons, also other continuous metrics for similarity could be used. For instance, one could use the same descriptor as in this Figure, but take Eq. (2.99) on p. 68 as a metric. Still, a pre-defined threshold usually is needed to decide whether these individuals are binned to the same niche, *i.e.*, if $d_{ij} \leq d_{\text{thresh}}$ between two individuals i and j .

The usual *greedy* algorithm for the niche-binning in OGOLEM takes place as follows: Find the *first* niche to which the new child can be binned by comparing it with each individual of the current pool. Here, such a loop starts at the best currently available individual and ends at the worst. If a *first* $d_{ij} \leq d_{\text{thresh}}$ is found, the new child just competes with the subset of individuals in this niche and can be inserted to the pool if it shows a lower (better) fitness than one of the individuals in the same niche. If no niche could be found, the new child can still be inserted to the pool if it is better than some other individual (*and* diverse with respect to other qualities).

Consequently, this single sweep does not calculate and “equilibrate” all niches in each such comparison as it would be the case in HC, which was explained in Section 2.6.2 on p. 65. The strategy here is essentially a greedy single linkage cluster method scaling as $O(N_{\text{GOCAT}})$ without calculating the full distance matrix, \mathbf{D} , and without undoing steps beforehand or re-clustering after successful additions. That is, if there were an even smaller distance between the new child later in the same loop, the child would nevertheless be assigned to the better (lower fitness) niche. The advantage of this procedure is that the sparse region in the high-dimensional search space, or in this case in the ESP-space, does not have to be pre-partitioned with a static grid, as used in many niching protocols for CSO in OGOLEM. The partitioning is rather only dependent on the appearing distances during the optimization.

This same niching, but with an eigenvalue-based CM descriptor for the comparisons, is investigated later in Chapter 5.

3.6 Quality Assessment of Catalysis

As we do a single-objective optimization,²¹ we have to define what we understand by a catalytic effect on a reaction by a surrounding GOCAT and subsume all those aspects into a single number, the fitness. This fitness should encode the utility as solution for the given optimization problem. Additionally, but dependent on the algorithmic framework used

²¹ Defining separate objectives and making a linear combination of those with defined weights is actually also already named multi-objective optimization in its simplest form. The weights and the fact of linear aggregation is already making the decision about how important each objective is, a decision not made otherwise when Pareto-frontiers are the last answer.^[29]

not always strictly necessary, some commensurability is advantageous, *i.e.*, that the actual value *difference* between two candidate solutions also mean something. Otherwise the fitness would just define an ordinal ranking. The final concrete number itself is usually of no meaning.²² In this Section, further details about the more sophisticated versions of the fitness function are discussed as they have been developed over time.

3.6.1 Static Fitness Function

A typical fitness function used in many **GOCAT** optimizations is shown in Algorithm 3.1 on the next page. Note that this is already the result of first rounds of **GOCAT** designs that lead to artifacts of *overfitting*. These are illustrated later in Section 3.6.2. Thus, successively more and more ingredients were added to the fitness function for the reactions tackled so far in order to come up with an increasing number of chemically meaningful candidate solutions. This **GOCAT** optimization problem can also be recast into an *implicit* regression problem where a continuous scalar function, the effective **PES**, is to be fitted by varying the constitution of a surrounding **GOCAT**. Implicit here means that mostly no known training values are available, *i.e.*, no explicit $\{\mathbf{R}_i, E(\mathbf{R}_i)\}_i^N$ pairs, but the final form such as monomodality, “vertical” stabilization with regard to the pristine gas phase reaction—*i.e.*, changed energy at the same molecular structures without changing the reaction path—and a small (single) barrier corresponding to a catalytic effect are the intended outcomes.²³

Algorithm 3.1 accumulates all types of different ingredients into one single objective quality measure:

- (1) Stationary points from the pristine reaction are (weakly) enforced by adding a penalty as fitness ingredient if $\|\nabla E_{\{\mathbf{R}, \text{TS}, \text{P}\}}\| > 10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ (usually), see Line 4 in Algorithm 3.1. After each calculation of a gradient norm for the two minima and the one maximum along the discretized **MEP**, the current fitness, f , is always compared with the current worst fitness in the pool, Line 7, called `immediateFallback(·)`, which is possible due to the *elitism* of our algorithm: If the incoming **GOCAT**, g , already had a worse fitness after this **SP** calculation(s), we can already return earlier, saving all further computations *without* any impact on the overall **GA** progression. This can be understood as a partial search space reduction technique, since in such cases the **GOCAT** fitness does not have to be evaluated completely.
- (2) The main catalytic barrier, $\Delta E^\ddagger = E_{\text{TS}} - E_{\text{R}}$, is heavily weighted in Lines 12–14.
- (3) Then, different (and also often varying) kinds of terms follow: the **TS** is to be stabilized, $E_{\text{TS}} \leq E_{\text{TS}}^{\text{ref}}$, where the latter is the known energy of the pristine **TS** without a **GOCAT**, checked in Line 18.

²²In the concrete problem, *e.g.*, optimizing some molecular structure in **CSO**, the fitness is identical to the energy, *i.e.*, it has a clear meaning. In this case, however, the lowest (negative) energy in the cluster is not known *a priori*, *i.e.*, the fitness has no lower bound. In parameter optimization, using a squared difference to a reference as fitness, it has a definite lower bound. However, this does not mean that this bound, zero, can actually be realized at all. Neither does this mean that a zero fitness would be *good* at all (*cf.* *overfitting*).

²³A **GOCAT** optimization in a supervised regression format emulating **COSMO**^[432] embeddings was also done and discussed in Ref. [433] of Chapter 6.

Algorithm 3.1: Typical static GOCAT fitness function.

Input: GOCAT: g

Data: weights/ingredients for penalties and which terms to include

Result: single fitness, f , representing multiple objectives

Function fitnessFunction(g):

```
/* initialize the fitness: all fitness terms are positive */
f ← 0
foreach i in { R, TS, P } do // check whether gradient norms are “ok”
4   ||∇Ei|| ← calcGradNorm(i)
   if ||∇Ei|| ≥ gradthresh then // gradient norm worse than threshold
       f ← f + weight · calcPenalty(||∇Ei||)
7       if immediateFallBack(f) then // fitness worse than worst g in pool
           | return fmax
       end if
   end if
end foreach
/* calc the main catalytic barrier */
12 ER ← calcEnergy(R)
13 ETS ← calcEnergy(TS)
14 f ← f + weight · (ETS – ER)
   if immediateFallBack(f) then // fitness worse than worst g in pool
       | return fmax
   end if
/* optional: enforce further restraints, e.g.: R must be stabilized */
18 if ETS > ETSref then // compare to energy w/o GOCAT
   | f ← f + weight · calcPenalty(ETS – ETSref)
   | if immediateFallBack(f) then // fitness worse than worst g in pool
       | return fmax
   | end if
end if
24 if ER > ERref then // compare to energy w/o GOCAT
   | f ← f + weight · calcPenalty(ER – ERref)
   | if immediateFallBack(f) then // fitness worse than worst g in pool
       | return fmax
   | end if
end if
/* calc (other) barriers and add those also to fitness */
30 f ← f + weight · checkMonomodality(calcEnergy({ 0, ..., N }))
/* check, whether the minima and maximum are at (or nearby) R, TS, P */
31 f ← f + weight · checkMaxMin(calcEnergy({ 0, ..., N }))
   return f (the aggregated fitness)
end
```

- (4) In principle, the barrier decrease can happen in the following three variations, by
- pure increase in energy of **R** (destabilization),
 - pure decrease of energy of **TS** (stabilization) or
 - a mixture of both the aforementioned.

Hence, in Line 24, an additional penalty can be added if the **R** frame is not stabilized. In some **GOCAT** optimizations all the above mechanisms are contained within the final pool. In some reactions, one can discriminate those and reach qualitatively different solutions.

- (5) Now, really all frames of the discretized **MEP** are evaluated by `calcEnergy({ 0, . . . , N })` and input to `checkMonomodality(·)` in Line 30. Here every single possible barrier is tracked, besides the main one, and appended also to the fitness by simply going through the list of energies.
- (6) Due to the gradient norm checks on $\{\mathbf{R}, \mathbf{TS}, \mathbf{P}\}$, the examined frames should be (close to) stationary points on the **PES**, but in loose **DOFs**, though, some (slightly) smaller energies as intermediate minima can be found (tracked in Item (5)). Even more stressing a meaningful path, the set, $\{\mathbf{R}, \mathbf{TS}, \mathbf{P}\}$, is checked to only contain proper minima and one maximum on the reaction profile (and not just on a flank to a lower minimum). This happens in Line 31. Thus, the frame position, i , of $\{\mathbf{R}, \mathbf{TS}, \mathbf{P}\}$ should not change (too much) and if it does, a *discrete* penalty is added that scales with the number of frames it is shifted.

The gradient norm penalty of Item (1) is crucial. Without a penalty of this kind, every point on an effective **PES** within the **GOCAT** could otherwise be sampled and considered as being sound, no matter how large the gradients were. Even equipotential reaction profiles could easily result without any chemical meaning at all.²⁴ Concerning Item (5), every intermediate barrier is recognized and treated as another $\Delta E_{ij} = E_j - E_i$ term added to the fitness, where $E_j > E_i$ for an intermediate minimum E_i and maximum E_j , assuming to have enough frames included in the path—which must be checked for the reaction at hand.²⁵

The penalties used are usually of quadratic type and the final weights are not given here. (Compare with the similar but less detailed explanations given in Chapter 6 for Ref. [433]; in the supporting information there, all weights are expatiated upon).

By just including, for instance, two frames, **R** and **TS**, or even just the **TS** alone as plain **TSS**, we could make the consternating observation of high intermediate energy barriers that are even higher than the single barrier in the perfectly relaxed pristine gas phase **NEB** path. This is one realization of overfitting. As long as such configurations are not looked at in the fitness calculation and hence such artifacts are simply overlooked during the **GA**, they

²⁴Just think of the non-convex Coulomb-singularity of a partial charge. With too much meaningless freedom, every energy as result could emerge, while the gradients could be arbitrarily strong.

²⁵So, in principle, one path with *one* barrier could compete with a path, *e.g.*, showing *two* barriers with half the height (if the weights are the same). Up to now, we have not observed such oscillating energy profiles in the converged pool which would not be chemically meaningful, as long as no further intermediate gradients are checked, too.

are detectable only by evaluating further frames on the reaction path (or other properties). Therefore, unless having a very simple short linear path and exposed vdW surface without much possible influence by the GOCAT, looking at more than just two frames is really needed, especially in a *global* optimization, where each lapse in setting up the fitness is ruthlessly exploited if this brings in a fitness benefit. Indeed, intermediately during the genetic algorithm (see below) we have seen multiple oscillations and intermediate barriers, also for, e.g., the DA reaction, which are penalized in this way.

The weights between all these terms as well as the mere presence of the latter are open for manipulation and define additional meta-parameters with all kinds of different GOCAT solutions at the end. Often, the impact of those can rarely be anticipated intuitively. Consequently, many trials with different settings are necessary. Also the internals of Line 30 and Line 31 could be changed (and maybe *should* be adapted) if in another (yet untreated) reaction further less meaningful GOCATs survive.

Thus, on the one hand, if there is not too much impact by a GOCAT possible as, for instance, by restricting them to small charge values/numbers and higher distances from the reaction frames that might follow a clear-cut electronic reorganization during the reaction and that might have no net charge, etc., one could try again to leave out the routines that have to calculate the full path (checkMonomodality(\cdot), checkMaxMin(\cdot)) for saving some computational resources. On the other hand, if especially very loose DOFs and longer reaction paths are tackled, over-stabilizations before reaching the actual TS might be exploited by the GA and thus exposed at the final candidate solutions; then, the routines are necessary or must be adapted to be even more restrictive.

In the current setting, this Algorithm 3.1 is exposed to a java code class that is dynamically re-compiled during the OGOLEM start-up such that this can be thought of as an “interpreted” programmable part.²⁶ Note also that due to maximal encapsulation, the SP calculation is delegated to the backend composed at runtime, e.g., to MOPAC, in order to calculate an energy and/or a gradient.²⁷ Using memoization, i.e., a simple cache, every subsequent call is of course not calculated *twice*, e.g., from Line 4 of R and Line 12, for example. After having calculated all energies of the whole MEP, cf. Line 30, no immediateFallback(\cdot) is needed anymore because the SP calculations are the most computationally expensive part.

3.6.2 Faulty Fitness Function

For further motivating the fitness function detailed in the last Section 3.6.1, we show results that could not be considered as being catalytic and that stem back to the very first steps within the GOCAT theme.

Using a very ancient fitness function, some impressions starting with not all, but just the most important frames are shown in Fig. 3.8 on the following page. At that time, it was still questionable whether only the TS alone should be included to start a plain TSS

²⁶ Usually, almost some type of a domain-specific language is assumed to be programmed to handle all kinds of very different properties in a fitness function under the free configuration by the user. For REAXFF fitting, the current author followed exactly that strategy. For GOCAT optimization, this could be amended in the future.

²⁷ Roughly, the (analytic) gradient calculation is maybe a factor of two more expensive than just an energy calculation. When not necessary, no gradients are therefore computed.

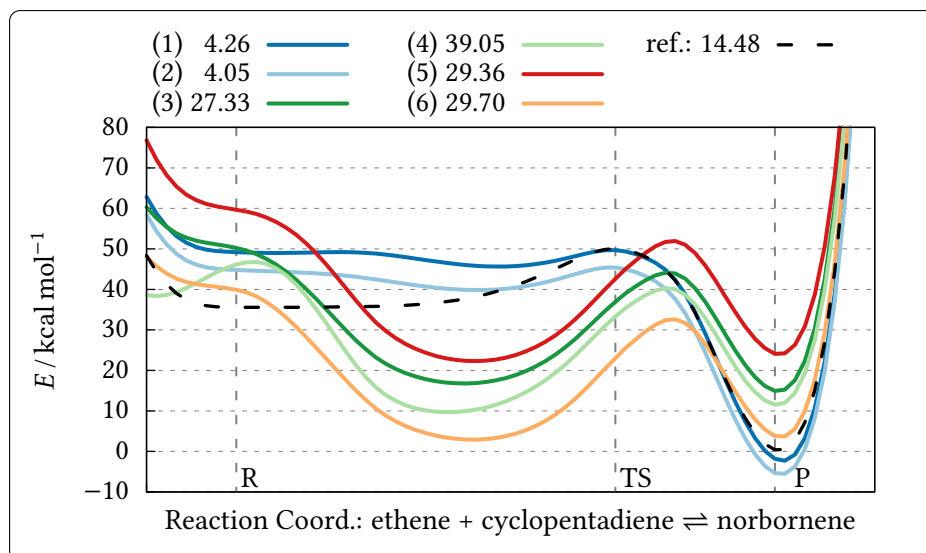


Fig. 3.8: Overfitting illustration using an ancient fitness function. The numbers given are the effective barriers, not those between the pristine gas phase stationary points that are indexed with **R**, **TS**, **P** and a dashed line. These profiles belong to a model of non-neutral **GOCATs** with $N_{\text{Ch}} = 10$ charges on a **vdW** surface of a **DA** reaction on **EVB-QMDF** level of theory.^[264] The separate **EVB-QMDF** potentials for this **DA** reaction are very similar to the original ones in Ref. [264, Fig. 2, p. 16717], *i.e.*, a slightly better fit than the ones of the Reference; these are also used in Section 6.3.1 on pp. 147ff. for the operator benchmark. See the main text for further discussions.

optimization. In this Figure, some hand-picked paths of actually different fitness functions that all *just* incorporate two or three frames of { **R**, **TS**, **P** } are illustrated. No gradient norms were checked at all and just the indexed frames (vertical grid lines) were used during the **GA** optimization. Afterwards, we can of course always look at the missing frames, and this result is plotted here. All settings have in common that the barrier is lowered, $\Delta E^\ddagger = E_{\text{TS}} - E_{\text{R}}$, that **R** might be destabilized and that also usually a relation of $\Delta E_{\text{TS}} < \Delta E_{\text{R}} < \Delta E_{\text{P}}$ holds with $\Delta E_i = E_i^{\text{GOCAT}} - E_i^{\text{ref}}$, *i.e.*, negative stabilization energies.²⁸ The single point we want to stress here is that just looking at the indicated single frames, we would misleadingly “see” a barrier of zero, *cf.* the cases (1)–(2) in Fig. 3.8, or even a negative barrier, *cf.* (3)–(6). Looking at the frames between the **R** and **TS** then reveals the non-chemical nature of the **GOCATs**. These show in-between overstabilizations with resulting *effective* barriers that are higher than the reference gas phase path. Yet, looking at (1) and (2) again, the barriers are quite small. Similar paths with some small overstabilizations that do not sum up to an effective barrier larger than the reference one, as in (3)–(6), would also be assessed as being fit in the (better) fitness function of Section 3.6.1. In contrast to the better fitness function, the essential difference is that gradients are not checked and also those are considerably too large at the so *pretended* stationary frames. Besides, intermediate overstabilizations

²⁸This was thought of a way of anticipating a product release: **P** is less stabilized than **R** and the **TS** must be stabilized most in order to have a catalytic barrier decrease. Of course, such and more restraints can always be implemented again if they are chemically meaningful for the problem at hand.

that possibly result in the large effective barriers are simply not noticed since no **MEP** with multiple frames is used for the fitness evaluation.

To conclude, gradient norms in this “vertical” setting have to be included, in principle. Dependent on the tackled reaction, one usually has to incorporate many frames, *i.e.*, mostly the whole reaction path in a sufficiently dense discretized form, in order to penalize such overstabilizations in-between. Often, being more conservative (restrictive) in this regard leads to better results, but for the question at hand, adaption is wise.

3.6.3 Adaptive Fitness Function

Finally, the generalization of the *static* or *vertical* mode of Algorithm 3.1 in Section 3.6.1 to a fitness function that fully relaxes the **MEP** surrounded by a **GOCAT** is discussed. This extended fitness function that uses the adaptive **NEB** implementation (*cf.* Section 2.5.1.3) is given in Algorithm 3.2 on p. 93. This delivers *non-vertical* or *adaptive GOCATs*. Here, first the usual fitness function, Algorithm 3.1, is used in Line 2 of Algorithm 3.2. If the fitness is in a catalytic region, *i.e.*, the final accumulated number is less than the one of the pristine gas phase path, the re-optimization of a **MEP** influenced by the **GOCAT** starts (Line 3). First, both end-frames are re-optimized to the nearest local minima (Line 4). Then, the sanity is checked (Line 5) and the chain of different sub-**NEBs** is applied (Lines 8–9f.) until the resolution of the new **MEP** is fine. When this holds, the **CI NEB** for optimizing the **TS** on the path is started in Line 16 and, finally, the “cosmetics” **NEB** is carried out in Line 20. This latter step does not only treat windows of the **MEP**, as done in the sub-**NEBs**, but it treats the whole **MEP** again with tight optimization thresholds. Each of these steps is separated by different sanity checks, *cf.* Lines 5, 13, 17 and 21. If all checks are fine, the final fitness of the re-optimized reaction path within the **GOCAT** follows in Line 24, and this new individual is returned in order to compete with the current population.

If anything goes wrong, the pristine path with the calculated fitness is returned; this path can either be exactly the same as in the static version or it can be an already new **MEP** of a successfully optimized **GOCAT** of an antecedent fitness function call. These checks mainly include examinations of the gradient norms and of the angles between the frames. The angles are defined by Eq. (2.92) on p. 60. In fact, convergence issues such as not reaching an acceptable threshold of the gradient norms at the new stationary points or angles that severely differ from being close to 0, pointing to loops or kinks, can appear and are sorted out in this way. These problems could, for instance, originate from non-appropriate meta-parameter settings including the energy–geometry resolution ratio ζ (*cf.* Eq. (2.87) on p. 58), maximal inter-frame path length allowed, the step-size control or from heavily non-linear paths such that the **CI** fails, besides others. For performance enhancement, these settings are adjusted to work well for the gas phase path without anticipating each imponderability within a **GOCAT**. In line with the mentioned imponderable affects, this could, instead, originate simply due to **GOCATs** that are not physical or meaningful embeddings. Especially, by using a **vdW** surface for the point charge entities, Coulomb implosion is an issue due to the non-convexity around the singularity of the charge center without any other repulsions

such as dispersion interactions of real atoms.²⁹ If all checks are fine, a new **MEP** with a distance below a pre-set threshold between each frame and enough energetic resolution is returned with its new fitness. Note that by using the gas phase path as reference and requiring the end-frames to be in a proper basin of attraction having not too large gradient norms before relaxation, in each such adaptive fitness function call, rather small changes are introduced and propagated to the population. In the next possible fitness function call, this path can again change such that after many cycles a completely new path can emerge.

Manifest follow-up questions regarding this scheme might be why not to introduce completely erratic big changes of the **MEP** in each call and how to guarantee the intended mechanism. Due to the relaxations, a former one-step process might evolve to a two-step process, for instance, of maybe even one step in a parallel reaction that had a higher barrier becomes the new minimum barrier and thus the catalytic bottleneck.³⁰

This Algorithm 3.2 at work and further discussions thereof are the main topic of Chapter 7.

²⁹That is, following the gradient direction can lead to an *increase* of the gradient and a decrease of the energy. Using **vdW** surfaces, separate atoms can be addressed explicitly and imposed to strong fields by the charges because of the small overall distances. Without compensating fields from other charges (symmetry) or by the other atoms (bonds), even stationary points of higher order are possible, *e.g.*, flanks/bifurcation points on the seam to the charge centers.

³⁰This goes in the direction of full reaction mechanism networks with, *e.g.*, needed graph theory to discriminate different reaction types by introducing such a discrete measure into the genotype of the **GOCAT** and ensuring the same **R** and **P** frames before and after the optimization. This is not implemented at the moment. Here, each **GOCAT** embeds *any* **MEP** with a fully converged path without any chemical configuration checks and competes with all the others of maybe different type during the **GA**.

Algorithm 3.2: Adaptive NEB fitness function for GOCAT design.

Input: GOCAT: g

Data: parameters for adaptive NEB & FIRE; weights/ingredients of for fitnessFunction(\cdot) (cf. Algorithm 3.1)

Result: GOCAT surrounding new MEP and its fitness

```
Function adaptiveFitnessFunction( $g$ ):  
    /* relax GOCAT and calculate fitness */  
2    $f \leftarrow$  fitnessFunction( $g$ )  
3   if  $f \leq$  threshold then // fitness in catalytic region  
4        $g' \leftarrow$  relaxEndPoint( $g$ ) // end-frames are locally optimized  
5       if sanityCheck( $g'$ ) not sane then //  $\|\nabla E\|$  small enough  
6           | return  $g$  (unchanged) with  $f$  as fitness  
7       end if  
8        $g' \leftarrow$  subNEB( $g'$ ) // non-linearly interpolated NEB between end-frames  
9       while resolutionCheck( $g'$ ) not fine do // check the geometric/energetic  
10          resolution, with an upper bound on the inter-frame distance at the end  
11          |  $g' \leftarrow$  addFrame( $g'$ ) // add one further frame via  $\zeta$  (cf. Eq. (2.87))  
12          |  $g' \leftarrow$  subNEB( $g'$ ) // window-based NEB of the new and some neighbor  
13          | frames  
14          end while  
15       if sanityCheck( $g'$ ) not sane then // no kinks &  $\|\nabla E\|$  small enough  
16          | return  $g$  (unchanged) with  $f$  as fitness  
17       end if  
18        $g' \leftarrow$  CINEB( $g$ ) // CI NEB for TS optimization  
19       if sanityCheck( $g'$ ) not sane then // see comment in Line 13 above  
20          | return  $g$  (unchanged) with  $f$  as fitness  
21       end if  
22        $g' \leftarrow$  TSFixedNEB( $g'$ ) // tight NEB optimization with fixed TS  
23       if sanityCheck( $g'$ ) not sane then // see comment in Line 13 above  
24          | return  $g$  (unchanged) with  $f$  as fitness  
25       end if  
26       /* now  $g'$  is fine: final fitness evaluation */  
27        $f' \leftarrow$  fitnessFunction( $g'$ )  
28       return  $g'$  (surrounding new MEP) with  $f'$  as fitness  
29   end if  
30   return  $g$  (unchanged) with  $f$  as fitness  
end
```

Publication: REAXFF Parameter Optimization

4.1 Scope of the Project

The aim of this project was the combination of REAXFF^[236,237] in form of its highly efficient implementation by the program sPUREMD^[410,411] with our EA-based global optimization suite OGOLEM^[390] in order to perform unbiased global REAXFF parameter optimizations, with all the established algorithmic benefits of OGOLEM. The meaning behind it was to provide an additional tool for all scientists when general or problem-specific (re-)parametrizations of the reactive force field, REAXFF, were needed for their own research without having to depend on other common less flexible approaches that often require expert insight. In this publication, regardless of the hard characteristics of this optimization problem, which is also strikingly shown, we demonstrate the performance reached both for the algorithmic progression with regard to the objective function minimization and for the general computational framework, including the parallel scaling. As one tool in the toolbox of researchers, this is supposed to provide one step towards the black-box optimization of REAXFF. However, actual chemical simulations with newly optimized FFs lie intentionally outside the scope of this publication. Hence, further tasks such as test validation to prevent overfitting and other practical issues such as training set creation and parameter selection are not addressed, which determines the whole optimization framing and its intricacy. These are tackled in other studies.^[138,247]

With respect to the context of this Thesis, we already use similar relational niching variants in the abstract parameter space in this publication as well. However, this is just one of many further improvements concerning this optimization class. Note that only the main implementations for the parallelization took place during the work for this Thesis, whereas the rest was described in much detail elsewhere.^[240] Thus, this opportunity is taken to address the topic of software engineering by illustrating one further design that had to be made and that is given as Complementary Information for this Chapter in Section 4.3.

4.2 Publication Data and Reprint

<i>Reference:</i>	M. DITTNER, J. MÜLLER, H. M. AKTULGA, B. HARTKE, EFFICIENT GLOBAL OPTIMIZATION OF REACTIVE FORCE-FIELD PARAMETERS, <i>J. Comput. Chem.</i> 2015 , <i>36</i> , 1550–1561, DOI: 10.1002/jcc.23966. ^[244]
<i>Submitted:</i>	March 12, 2015.
<i>Accepted:</i>	May 15, 2015.
<i>Contribution:</i>	During the Master’s Thesis ^[240] of the current author: Implementation of REAXFF optimization in OGOLEM including also many algorithmic improvements for metric-space-based real vector optimization that are discussed at length in Ref. [240] and shown in the first part of the paper (including “SiOH benchmark”). During the present Thesis: Implementation of parallelization improvements for shared memory computing (Section “Scaling” on p. 106). Major contribution to analyses, discussions and the text.
<i>Graphic:</i>	Illustrated in Fig. 4.1.
<i>ESI:</i>	Printed on pp. 288–293 in Appendix B.
<i>Copyright:</i>	Reproduced with permission from “Journal of Computational Chemistry”. Copyright 2015 Wiley Periodicals, Inc.

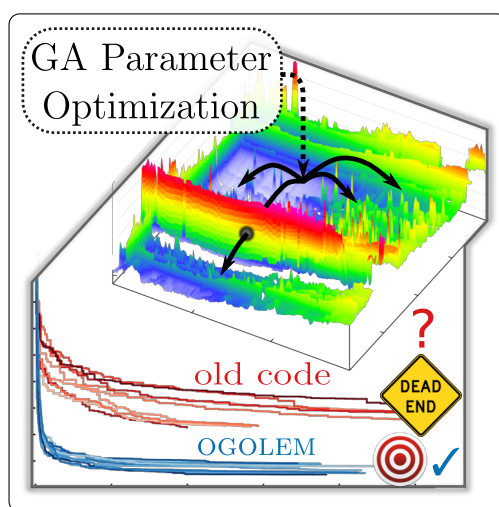


Fig. 4.1: Original “Table of Contents” graphic. Note that the 3D fitness surface shown here really belongs to this optimization problem—but was reused for illustration purposes also in other contexts.

Efficient Global Optimization of Reactive Force-Field Parameters

Mark Dittner,^[a] Julian Müller,^[a] Hasan Metin Aktulga,^[b,c] and Bernd Hartke*^[a]

Reactive force fields make low-cost simulations of chemical reactions possible. However, optimizing them for a given chemical system is difficult and time-consuming. We present a high-performance implementation of global force-field parameter optimization, which delivers parameter sets of the same quality with much less effort and in far less time than before,

and also offers excellent parallel scaling. We demonstrate these features with example applications targeting the ReaxFF force field. © 2015 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23966

Introduction

Chemical reactions can be simulated with convenient degrees of accuracy and generality by classical-mechanical molecular dynamics for the nuclei, with on-the-fly calculation of the internuclei forces via quantum-chemical methods.^[1,2] However, even with present-day high-performance computing resources, only system sizes of 100–1000 atoms are accessible, and one week of computing yields only 2–200 ps of simulated time. This is to be contrasted with classical-mechanical molecular dynamics using typical biochemistry force fields. Then, using the same hardware and similar computing times, much longer time scales are accessible (high-end simulations of explicit protein folding have propagated 17,500 atoms for 8 ms^[3]), as well as much larger systems (up to 134 billion atoms^[4]).

Due to the use of fixed atom types and nondissociative harmonic oscillators, however, force fields of this kind cannot be used to simulate chemical reactions where covalent bonds are broken or formed. This gap can be bridged by separating a big system into a small quantum-mechanical (QM) and a larger molecular-mechanical (MM) part, which also requires the introduction of suitable models for the boundary between the two parts. Such QM/MM approaches^[5–7] are widely used despite some of their shortcomings: Besides requiring a careful treatment of the QM/MM-boundary, its very predefinition prescribes where reactions can or cannot occur—but this knowledge may simply not be available *a priori*. Last but not least, the QM and MM parts have to evolve in synchrony, therefore, the performance of the QM part often limits the overall performance.

These problems can be circumvented with reactive force fields, either by using them on their own or in combination with a QM/MM approach. Currently, reactive force fields are developing from isolated niches toward broader ranges of application, and several different approaches have been proposed.^[8] Besides reactive force fields specialized to particular groups of elements,^[9–11] and recipes for combining particular reactants and products,^[12–15] there are also reactive force fields that aim at general applicability. Two of these, COMB

and ReaxFF (see Ref. [16] for a combined review), have gained considerable popularity in computational materials science and computational chemistry, respectively.

For high accuracy, force fields need to be fitted to a reference data set through a parameter optimization procedure. Well-founded methodologies for assembling reference data sets are largely lacking for this frequently needed procedure.^[17] Due to the large number of parameters to be optimized and the nonconvex nature of the search space, multistart techniques based on local optimization algorithms are problematic.^[18] Therefore, nondeterministic global optimization strategies, for example Genetic Algorithms (GA),^[19] have been used by several authors.^[20–30]

The parameter optimization problem for reactive force fields is harder than that of traditional force fields, because there are far more parameters per atom, these parameters are more strongly coupled, a significantly larger reference data set is needed, and we have limited knowledge about the relationship between reference data items and force field parameters. GA methods have successfully been applied to this challenging task,^[31–33] including a GA optimization study of ReaxFF parameters for SiOH^[34] and azobenzene^[35] by one of the present authors.

The techniques used in Ref. [31–35] are single-objective GA optimization techniques. Recently, a number of studies using multi-objective GA techniques to optimize ReaxFF parameters were published^[36–38] (see Section Related Work). In a single-objective scheme, it is necessary to predetermine the weights for individual entities in the fitness function. There are no such

[a] M. Dittner, J. Müller, B. Hartke
Institute for Physical Chemistry, Christian-Albrechts-University,
Olshausenstr. 40, 24098 Kiel, Germany
E-mail: hartke@pctc.uni-kiel.de

[b] H. M. Aktulga
Department of Computer Science and Engineering, Michigan State
University, East Lansing, Michigan 48824

[c] H. M. Aktulga
Computational Research Division, Lawrence Berkeley National Laboratory,
Berkeley, California 94720

© 2015 Wiley Periodicals, Inc.

requirements in multi-objective methods as they optimize multiple objective functions simultaneously. However, this attractive property of multi-objective methods comes at the computational expense of increased population and search space sizes during the search. Also the user is left with the task of post-selection of suitable candidates from a (possibly very large) number of Pareto optimal solutions (cf. Section Related Work). Hence, in this work we continue using the single-objective paradigm.

Force-field fitting in practice is an iterative process, repeating the following steps until convergence: "A: definition of the optimization problem" (choice of training set entries, selection of force-field parameters to optimize, etc.), "B: optimization of force-field parameters," and "C: tests of the newly optimized force fields, within the training set and outside of it". All of these steps are challenging and in strong need of further method development. In this work, we have focused on improving step B, leaving steps A and C for future work. Of course, improvements in B will directly benefit also steps A and C.

In this article, we present further progress in algorithms and implementation to our earlier work on a single-objective GA optimization framework for ReaxFF.^[34,35] We combine sPuReMD,^[39] an advanced implementation of ReaxFF, with OGOLEM,^[40,41] an advanced general evolutionary algorithm (EA) optimization suite. We show that the resulting framework produces results of at least the same quality as with our previous setup,^[34,35] but in significantly shorter real times, offers better scalability and provides better user support and accessibility. In Section Methods and Techniques, we briefly summarize key features of both OGOLEM and sPuReMD and discuss their combination. Section Results and Discussion presents comparisons between our earlier program suite^[34,35] and the present one. Related work on GA-based optimization of ReaxFF parameters and the distinguishing aspects of this study are discussed in Section Related Work.

Methods and Techniques

Background information: OGOLEM

OGOLEM is an object-oriented, easily extensible, platform-independent global optimization framework based on EAs, especially in the realization of GAs.^[40,41] It combines thread-level and MPI-level parallelism to achieve high scalability on shared memory as well as distributed memory architectures. The OGOLEM framework embodies our accumulated knowledge on nondeterministic global optimization in general and on EAs in particular,^[42,43] for various applications: cluster structures,^[44–54] protein folding,^[55] potential fitting,^[34,35,56–60] molecular design,^[61] and abstract benchmarks.^[62]

EAs^[19] borrow nomenclature from natural selection and evolution processes. To treat manifold optimization problems in a problem-independent manner, the problem specific system information, that is, everything that is defined as (indirect) input to the optimization function, is encoded as a genotype, a possible solution candidate is called an individual and the

set of all individuals (and therefore their genotypes) present at a certain point in time is dubbed the genetic pool. The genetic pool is refined iteratively through genetic operations: Crossover causes exchange of genetic material between two individuals and mutation changes the genotype of a single individual. For these operations, individuals are typically selected by a combination of random choice and preference for the currently best (fittest) individuals. By repeating this selection and modification process, better individuals found at each round replace older ones. Assuming enough resources, this process would eventually yield the globally best individual. Obviously, the evolution of individuals in a genetic pool can be performed simultaneously, making it straightforward to parallelize an EA.

The global optimization power of EAs goes beyond the possibilities in a natural evolution setting. In natural evolution, there is no need to find a global optimum; for any species or individual, it suffices to be better than their geographic neighbors and logistic competitors. Instead, EAs are good for global optimization because via crossover (a) they can exploit (partial) separability of the optimization problem even in the absence of any explicit knowledge about its presence and (b) they are able to make long-range "jumps" in search space. Due to the continuous presence of multiple individuals that have survived several selection rounds, (c) it is ensured that these "jumps," based on information interchange between individuals, have a high probability of landing at new, promising locations. Last but not least, by admitting operators other than the classic crossover and mutation steps, (d) it is possible to extend EAs within this abstract meta-heuristic framework^[19] with nice features of other global optimization strategies, too.

EAs are especially valuable when dealing with challenging and time-critical optimization problems. The straightforward parallelism and intrinsic high scalability property of EAs provide an advantage over other strategies which are either serial by nature or where parallelization facilitates decoupled or only loosely coupled task-level parallelism. EAs constantly share a common knowledge among workers while still exhibiting excellent scalability^[40] through extensions developed in our group.^[51]

OGOLEM can interface with different backend software for the computation of properties such as energies, gradients, or frequencies, and focuses on providing the best high-level optimization algorithms and task management strategies. The external codes in turn are expected to provide the best possible implementation of their task. However, due to the algorithm detailed above, EAs are not limited by the scalability of the underlying property evaluation, allowing for the best possible implementation with only limited scalability concerns for the external code (cf. Section Scaling). The general GA-iteration cycle of OGOLEM together with some algorithmic options and backends is illustrated in Figure 1.

Extensions to OGOLEM

In this section, we present extensions to the OGOLEM framework consisting of newly added utilities to provide support for a

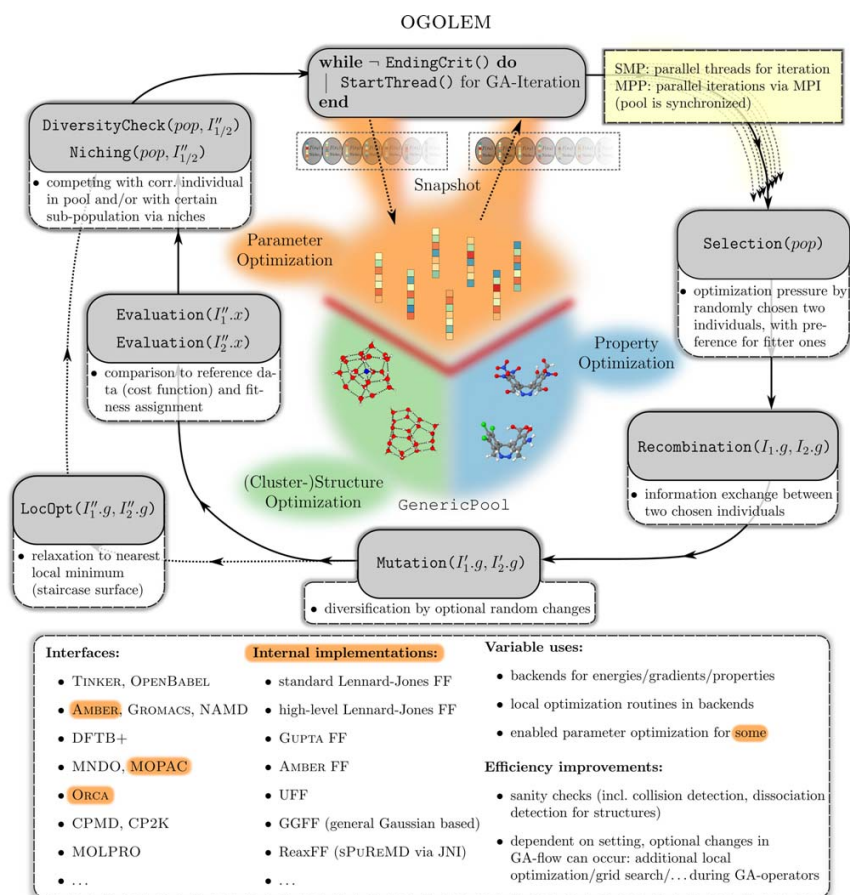


Figure 1. Flowchart of GA-iterations within OGOLEM (clock-wise). Certain important GA-operator steps together with their abstract role are shown. Many different implementations for these single operators are available, like described in the main text and also published in the cited literature of this Section Background information: OGOLEM. I stands for individual, g for genotype, and x for phenotype. Between the generic tasks, the work-flow might vary slightly (e.g., local optimization and additional algorithmic checks) as this figure mainly describes the parameter optimization task (orange color). Further details on topics like the alternative local optimization engine and niches are described in the main text. "Snapshot" refers to our generation-free GA-pool algorithm, found in Ref. [51]. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

wide variety of training data, and implementations of new genetic algorithm ingredients for high quality parameter optimizations.

Training set. In fitting of force field parameters to reference data, practical requirements are very different depending on whether the model is a simple function like the Lennard-Jones potential or a more elaborate one like ReaxFF. In the latter case, several kinds of training data (e.g., molecular properties) need to be included in the reference data set. The training data are also linked to geometric data (e.g., molecular structures). Thus, we extended the OGOLEM framework to enable support for different kinds of data. As a result, different molecular properties, such as absolute energies or difference energies with arbitrary prefactors (i.e., reaction equations), gradient information, partial charges, heat of formation, dipole moments, as well as geometric information (e.g., bonds, valence angles and dihedral angles) and cell parameters (for

arbitrary periodic crystal structures) can now be used as reference data in OGOLEM. Even a seemingly exotic property for a force field, an ("electronic") excitation energy, was implemented with a consistent treatment of multiple force fields at the same time; extensions to its first use^[35] will be the topic of future publications.

An entry in the reference data set can be evaluated either through a "single point" computation (i.e., directly using the molecular structure provided as input), or after performing a local geometry optimization first. Clearly, the latter is necessary for all geometric data (bond distances, angles, and dihedral angles). For other items in the reference data set, the user can choose to carry out a single-point computation directly or performing a local geometry optimization first. For local geometry optimizations, one of the several routines already available in OGOLEM can be adopted using the current parameter values of the GA individual to be evaluated. It is also possible to include diverse restraints into these local geometry optimizations.

Additional classes and utilities were added to `OGOLEM` to establish a general input structure that can handle the necessary information: (1) a “template” force field file, providing fixed values for parameters not included in the optimization (allowing each individual to represent a full set of force field parameters), (2) a parameter definition file, specifying the parameters to be optimized and their value ranges, (3) a training set file, containing reference data from higher-level computations and/or experiments, which also specifies the relative weights of the data items in the objective function, and (4) a geometric information file, containing the different molecular structures (atoms/molecules/crystals) which the entries of the training set are linked to. These extensions to `OGOLEM` were designed to retain compatibility with the corresponding input files of the original `ReaxFF` implementation by Adri van Duin et al.,^[16,63] where a nonglobal strategy of successive one-parameter parabolic extrapolation^[64,65] was used for parameter fitting. This enables us to reuse older input settings for our previous GA-implementation^[34,35] without changes, and to easily compare this older implementation with the present one, which is one aim of the present article.

As in earlier work,^[34,63,65] the objective function of the optimization procedure is defined as the aggregated sum of quadratic differences (“error sum”) for each molecular property given in the training set (further information—also with respect to the `RSSR` case below in Section Disulfide application example—can be found in the supplementary information of this publication). At this point, further savings are enabled via a “smart” training set evaluation: Our new `OGOLEM` extensions use caching techniques to remember already calculated items and avoid unnecessary recalculations within an iteration. Thus, only those properties that are actually needed are calculated once, in contrast to older implementations where redundant properties were calculated for almost every item in the geometry input file. `OGOLEM` also interprets the training set in order to recognize larger blocks of difference energies: For example, a dissociation curve that is specified as a contiguous block in the training set automatically leads to the creation of a “reference energy” for all energies in this block, preventing some redundant overhead and object creations. Because of our evaluation of the complete training set (i.e., parallelization at GA iteration level, cf. Section The `ogolem-sPuReMD` combination) as a serial aggregated sum we are now able to stop the fitness evaluation of one GA individual before all contributions to this sum have been calculated, when synchronous sanity checks show that the partial error sum is already larger than that of the worst individual in the current pool. In such a case, there is no chance for the new GA individual to be added to the pool after completion of the error sum calculation. This feature was dubbed `ImmediateFallBack`. Since this feature anticipates the result one would get without this feature, it does not change the development of the GA pool but saves computer time. This can also be understood as a partial on-the-fly search space reduction technique.

General parameter optimization GA-algorithms. Further extensions were made to `OGOLEM`, introducing several new crossover

and mutation operators as well as new niching techniques. Now the full range of possible crossover operators is available in `OGOLEM`, from single-point through two-point and k -point up to uniform crossover. We have added arithmetic crossover operators that not just swap but mix certain genes of the genotypes. For instance, different genes (parameters) of the elders are mixed as a randomized mean (randomized weights for father and mother) so that intermediate parameter values arise for the children.^[19] This is especially useful for mixing and creating new parameter values at the end of a GA run, when similar individuals dominate. For mutation, there are nullary to unary operators, that is, mutation as a partially random reinitialization within the parameter boundaries or locally around the current parameter values. To our experience, a stronger exchange between individuals often accelerates the optimization procedure. Hence, not just a single point crossover, but a k -point crossover involving up to 20–30% of the optimizable parameters should be chosen. Also, a mix of reinitialization mutation (nullary) and unary mutation as local “hillclimbing” has been found useful, particularly in the later stages of GA-runs. As standard feature of `OGOLEM`, any desired (weighted) combination of these (and more) operators and protocols can be chosen by the user. Also available are hybrid local optimization routines, allowing for further relaxation of the individuals to the nearest local minima, via local hillclimbing or local gradient-free optimization. This can be applied during the global optimization, after preliminary iterations, or as a-posteriori refinement (restart with a seed of old pool). Former and current experience shows that these additional local optimizations can be beneficial at the very final stages of the optimization or as post-processing to reach the best local optimum of selected individuals. For more general usage, local optimizations are too expensive, since no analytic gradient is available. Also, the ruggedness of the search space (as shown in Section “Objective function surface”) may render local optimization inefficient in initial stages of the GA.

As in previous work,^[45] we employ niching^[19] to maintain diversity within the population and to decelerate premature convergence. We have implemented different variants of niching, based for example on grid-mapping. In one version, the floating-point values of every parameter in a genotype are mapped onto a population vector of integers, leading to a coarsened representation of all individuals. The integer number of identical or different genes then serves as a common identity measure. Alternatively, vector norms between the genotypes are used to enforce a minimal distance between genes of different genotypes. In these two cases and in some of the others, the niches are based on a relational measure of a snapshot of the current population, that is, they are largely transient instead of predefined.

Moreover, all new implementations and the code-basis of `OGOLEM` pay attention to user-friendly, keyword-based control of input and output with a policy to check for input inconsistencies. For example, upon a missing geometry entry or a simple typo, the user is informed and the calculation does not start. Thus, only a small and clearly arranged input is necessary, and just a small amount of I/O takes place (using mainly binary

serializations of objects), reducing redundant overhead to a minimum. Instead, after the calculation, the desired information is read from the binary pool and written to disk.

Background information: sPuReMD

As described above, OGOLEM is ideally suited for the challenging task of globally optimizing parameters in reactive force fields. A crucial component in our framework is an efficient implementation of the target force field, in this case ReaxFF. For this purpose, we use the sPuReMD open-source software.

sPuReMD (serial Purdue Reactive Molecular Dynamics program)^[39] is an optimized implementation of the Reax force field (ReaxFF).^[63] sPuReMD uses novel algorithms and data structures to achieve high performance in force computations while retaining a small memory footprint. An optimized binning-based neighbor generation method, elimination of the bond order derivatives list in bonded interactions, lookup tables to accelerate non-bonded interaction computations and a preconditioned GMRES solver for the charge equilibration (QEq) problem^[66] are the major algorithmic innovations in sPuReMD.^[39] The dynamic nature of the bond, 3-body and 4-body interactions in a reactive molecular system presents challenges in terms of memory management and data structures for efficiently computing bonded interactions. sPuReMD introduces novel data structures to store 3-body and 4-body interactions in a compact form. Its dynamic memory management system automatically adapts to the needs of input system over the course of a simulation. The dynamic memory management capability significantly reduces the overall memory footprint and minimizes the effort to setup a simulation. sPuReMD has been shown to outperform the LAMMPS/REAX package by a factor of 6–7 on various systems while using only a fraction of the memory space.^[39]

PuReMD, a distributed memory code with MPI-based parallelism, has been developed based on sPuReMD to enable the study of large molecular systems.^[67] PuReMD has been ported into LAMMPS software suite as the USER-REAXC package. PuReMD and USER-REAXC have been used by researchers around the world to study phenomena ranging from water-silica surface interactions^[68] to oxidative stress in lipid molecules.^[69] Recently, Kylasa et al. have developed the GPU accelerated version of the PuReMD codebase (Kylasa et al., in preparation).^[70] The entire PuReMD codebase is freely available with GNU Public Licence on the web.^[71]

The ogolem-sPuReMD combination

We combined OGOLEM with sPuReMD rather than with PuReMD: As mentioned, the latter includes MPI-parallelization and is aimed at MD for large systems. In our target setup, however, we mainly need ReaxFF single-point evaluations or local geometry optimizations of small systems, as OGOLEM backend. For these tasks, parallelization of ReaxFF incurs more overheads than benefits, and it would make the whole setup more difficult to handle. As discussed in Ref. [34], parallelization at two other levels are possible: across reference items and across GA individuals. Previously, we had chosen the former option.^[34,35] Here, we choose the latter, since OGOLEM is already equipped

with excellent parallelization at the GA level. One could argue that it is better to parallelize at this level since there the needed communication is minimal by construction, leading to better scalability. However, in both implementations there still is the possibility to also parallelize at the respective other level. We leave this option for future work.

Most of the core code of OGOLEM is already formulated not only object-oriented but also generically, that is, for most operations it does not matter if they are applied to cluster structures or to parameters in a fitting problem or to other items to be optimized. In this form, OGOLEM was already used and validated for many of the optimization problems mentioned in Section Background information: OGOLEM. This greatly facilitated the task of merging OGOLEM with the ReaxFF-backend sPuReMD to allow for the global optimization of ReaxFF parameters. Nevertheless, several decisive extensions had to be made, which are described below.

Backend for ReaxFF calculations. sPuReMD (implemented in C) was slightly changed and is now embedded as native code into OGOLEM (implemented in Java) as a dynamic library. To this end, communications between C- and Java-code via Java Native Interface (JNI)^[72] and modifications to sPuReMD were implemented. Hence, no further I/O operations are necessary, as OGOLEM manages the complete optimization flow (globally and locally) and the training set. Whenever a ReaxFF-evaluation for items like energies, gradients or charges is needed, the corresponding items (geometries, and current force field parameter values) are passed to sPuReMD. The main features of the latter are identical to the ones described in Section Background information: sPuReMD. However, to make these calls via JNI efficient and scalable, an extended new memory management scheme was implemented on top of the existing one in sPuReMD: A new thread-safe address space handling was implemented into OGOLEM, which passes also a starting address to sPuReMD within every call. On this starting address, a complete “scratch” space is built for all sPuReMD-specific simulation variables (`structs`) in sPuReMD that is still dynamically handled and is changed to the current needs (small footprint). Additionally, after some initial calls, that is, during first training set calculations of the first GA-iteration, an upper bound of memory per address space is determined to treat all items, including the biggest one incurred by the combined geometry and parameter set input. This allocation survives ensuing returns from the C-code (sPuReMD) to the Java code (OGOLEM), because its leading address is given back to OGOLEM where it is further managed and reused in subsequent calls without any further concurrency locks or similar problems. This saves almost all of the later memory allocations and deallocations and provides us with further absolute timing and concurrency scaling benefits.

Results and Discussion

SiOH benchmark

As a benchmark of the new OGOLEM/sPuReMD-combo, we revisit the optimization task of a previous publication.^[34] A search

space consisting of 67 parameters is defined, and the training set is based on 304 chemical geometries containing Si, O, and H atoms. Many local geometry optimizations with multiple restraints are needed for calculating certain training set items. Periodic crystal structures are involved, some of which also require optimization of the crystal cell. Finally, also some single-point calculations for different energy entries and a few charge properties occur, involving the main charge parameters that are used for charge equilibration (further details including the complete training set can be found in the Supporting Information of Ref. [34]). This training set had been established and used by the van Duin group before,^[68,73] employing their own nonglobal, iterative parameter optimization method.^[64,65] In our previous publication,^[34] we had shown that our old GA/ADF setup could already improve upon the van Duin results, despite the complete absence of domain-specific knowledge and experience on our part. However, this still needed several series of many program runs and elaborate sequences of parameter range tunings. Here, we demonstrate that our new OGOLEM/sPuReMD-combo simplifies and accelerates this task considerably through its advanced features.

Objective function surface. First of all, to stress why elaborate nondeterministic algorithms are indeed necessary, we present typical views of the search landscape in Figures 2 and 3. They show the objective function values, that is, a “fitness landscape,” in two-dimensional subspaces of the hyperdimensional search space. As the objective function is mainly a quadratic difference function between reference values and calculated ReaxFF values, a smaller value can directly and metaphorically be seen as a better fitness of that individual. Therefore, Figures 2 and 3 can also be interpreted as a systematic scan across 22.5×10^3 possible individuals each. This illustrates that the total 67-dimensional search space can of course not be scanned entirely (in fact, it grows exponentially with dimensionality), which is one reason for our use of nondeterministic algorithms. A second reason is that besides its astronomical size also the structure of the search space is challengingly complex. At least in the initial stages of the search, situations as the one illustrated by Figure 2 are to be expected: Many landscape features are clearly visible that are signatures for difficult optimization problems, for example, epistasis (not symmetrical due to parameter correlations), ruggedness, deceptiveness (misleading gradient information), and of course multimodality, as many different minima-regimes can be seen.^[19] Therefore, with local gradient-following algorithms, several restarts are needed to overcome this complexity. Such a strategy can only succeed for small dimensional problems in practice due to the exponential increase in the number of restarts necessary.

Figure 3 depicts the landscape for the same two parameters within the same boundaries as in Figure 2, except that an individual from a late stage of optimization was taken as basis for the remaining 65 parameters. Clearly, the landscape looks very different now. This situation demonstrates that there are significant correlations between parameters to be optimized, yet another feature that makes optimization difficult.

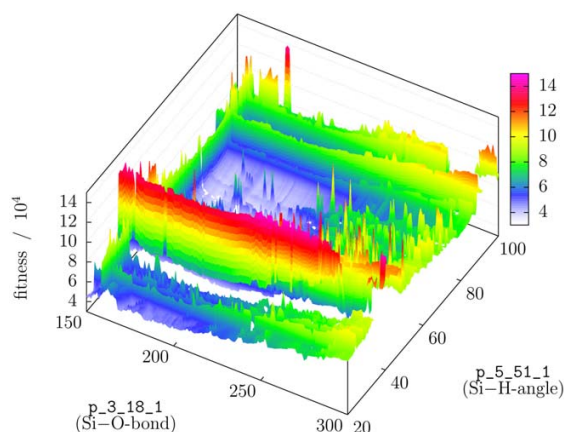


Figure 2. 2D objective function surface for two parameters (out of 67) of the SiOH training set. An intermediate solution with an error sum of about 100,000 is shown, occurring during a GA-run. Some interpolation due to smooth color progressions is implied. Transparent regions are erratic or mountain-like “pillars” with objective function values larger by several orders of magnitude; they are made transparent for clarity. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Given these difficulties, one may wonder how it was possible to arrive at useful ReaxFF parameters with locally optimizing methods.^[16,63–65] We suspect that this is largely due to two factors: (1) experience (domain-specific knowledge), which can enter in various ways, for example via selection of suitable starting points for multistart local optimization or (perhaps even more importantly) via restrictions on search space size (parameter variation limits) and dimensionality (selection of parameters to optimize); and (2) simplification of the search landscape in the vicinity of good solutions. The latter feature is strikingly illustrated by again comparing Figure 3 to Figure 2.

These difficulties and computational complexities of the parameter optimization task can be addressed better using

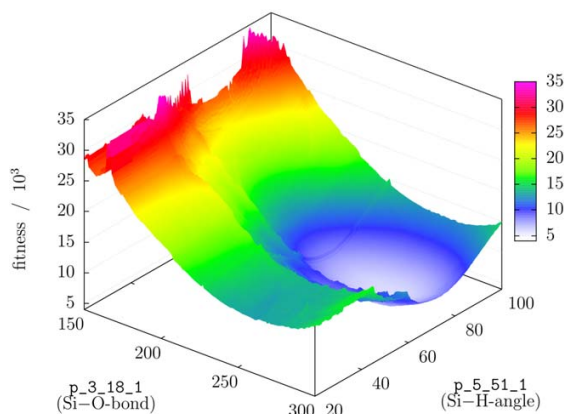


Figure 3. 2D objective function surface of the same two parameters as in Figure 2, but for a good solution near the end of a GA-run, with an error sum of 6150 (close to the global minimum). Again, some interpolation due to smooth color progressions is implied. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

nondeterministic metaheuristic GAs with more than exponential optimization progression from Figure 2, mostly randomized GA-starting individuals, to Figure 3, as we will show below. This way we partially substitute human expertise with computer power.

Comparison of former and recent optimizations. The typical optimization progress in a high-dimensional search space for ReaxFF parameter optimization is shown in Figure 4. After random initialization of the population (for instance, 200 parameter vectors are created to start each calculation), the objective value of the best individual in the current population decreases in faster than an exponential progress initially (in this case, up to about 2×10^3 iterations). This rapid initial fall-off has two causes: The ease with which the initial random seeds can be improved upon, and the information exchange at the beginning of the GA, leading directly to even more promising regions of the search space and establishing different promising 67-dimensional parameter vectors. Then, a slower progress takes over (looking almost like a “plateau” when compared to the initial phase), mainly because it becomes harder to further improve upon the already good solutions present. Finally, progress becomes slower than practically useful, which is dubbed “premature convergence.” The aim is to find the global minimum before this happens.

This general GA behavior is clearly visible in all curves displayed in Figure 4. However, there is a clear difference between the behavior of the old and the new implementations: The level of the plateau reached in the later stages of the GA is significantly lower in the new implementation. As a result of the improvements described in Section Methods and Techniques, we are now able to reach a mean fitness of about 4900 after 20×10^3 iterations (Fig. 4). Representative and comparable runs of the same length with our older codebase only lead to a fitness value of about 14,300. Thus, the solutions at this stage are improved by a factor of almost 3.

This quantitative improvement is likely to lead to qualitative changes. Figure 4 also shows a comparison with the error sum of 6646 that was found using nonglobal, iterative procedures for the same SiOH case.^[68,73] (Note that exact values of the error sum depend on some technical details such as convergence thresholds of local geometry optimizations, distance cutoffs, etc. The value of 6646 quoted here is obtained under present settings that are slightly different than those in Ref. [34], where the reported value was 6455). Runs with our new `OGOLEM/sPuReMD-combo` drop below this mark already within the first few thousand steps. In contrast, using the older codebase and within single runs of the given total length, we cannot reach the value of 6646 achieved by a non-global, iterative procedure.

In earlier work,^[34] this prompted us to do further series of runs, starting from the best individuals reached so far and also shrinking the parameter search space based on the parameter variations observed in the previous round. Additionally, we topped off this procedure by local, derivative-free parameter optimizations to get more quickly to the true bottoms of the wells found by the GA. This way, we previously managed to

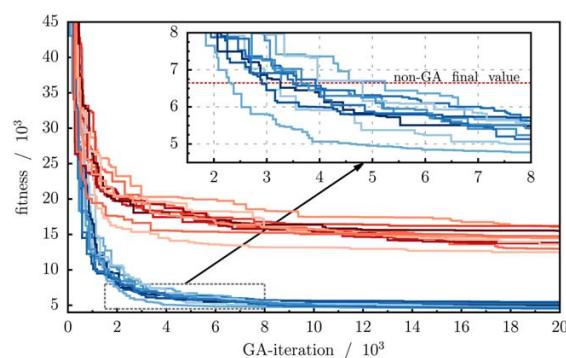


Figure 4. 10 selected but representative GA-runs with our old code (red lines)^[34,35] for the SiOH benchmark, in comparison to 10 with our new `OGOLEM/sPuReMD-combo` (blue lines, present work). The objective function value (i.e., error sum that is used as fitness for our GA) for the best individual of the current population in each run is plotted against the GA-iteration number. The magnification inset also includes the originally published best error sum for this SiOH case (horizontal line marked “non-GA final value”). It is easily surpassed by our new GA setup within a few thousand steps. As published before, runs with our old setup also eventually dropped below this mark, but only after considerably more time and effort. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

improve upon the 6646 mark, still without employing any domain-specific knowledge. However, the overall procedure required considerable user effort and far more computer time than the runs shown in Figure 4. The present `OGOLEM/sPuReMD` results obtained could also be improved further by performing additional runs seeded with promising individuals from former runs and by shrinking the search space. Leveraging the computational efficiency of our new codebase, the GA itself could be further improved by extending the pool size and the niching tightness, or by increasing the number of iterations. Or, putting it differently, with the new `OGOLEM/sPuReMD-combo` we can now reach far better individuals than before within just a single run for this SiOH benchmark, thus further reducing the need for additional work and cleverness on behalf of the user.

This last statement is illustrated in Table 1, where we provide the mean fitness values together with absolute mean wall-clock timings for different variants of our codebase. As shown in this table, leveraging the linear scaling property (Section Scaling), we are now able to use large numbers of cores efficiently (compare columns `OGOLEM2` with `OGOLEM2-p`). Thus, we significantly shorten the wall-clock time needed to reach good solutions. Better force fields with objective function values lower than the literature value of 6646 are identified within a few hours (4.2 h in the last column of the table), while no single runs of the older code could come close during the entire execution time of 142.2 h. Lower objective scores could already be obtained using the initial version of our `OGOLEM/sPuReMD` code (cf. column `OGOLEM1`). Our performance optimization work described in Section The `ogolem-sPuReMD` combination yields significant speedups in `OGOLEM2` when compared to `OGOLEM1` (134.0 h vs. 58.4 h in this example). This optimization included the utilization of the memory scratch-space, a simplified grid-space mapping of atoms for

Table 1. Comparison of absolute (wall-clock) timing results and the objective value reached with different GA implementations, averaged over 10 runs of 20×10^3 iterations each.

	Old code ^{[a],[b]}	OGOLEM1 ^{[b],[c]}	OGOLEM2 ^[b]	OGOLEM2-p. ^[d]	OGOLEM2-p. < 6.6 k ^[e]
Fitness	14,335(1242)	5061(581)	4833(297)	4717(347)	6468(257)
Timing (h)	142.2(16.0)	134.0(21.7)	58.4(7.9)	15.5(1.2)	4.2(1.4)
Iterations ^[f]	20,000	20,000	20,000	20,000	4042(1413)

Standard deviations are given in parentheses. [a] GA/ADF implementation of Ref. [34]. [b] 10 cores (threads) in parallel on 4×AMD Optreron 6274 16-Core, 2.2 GHz with 32×DDR3 PC1333 Reg. ECC. [c] First implementation without performance enhancements. [d] Same as OGOLEM2, but with 40 cores in parallel. [e] Same as OGOLEM2, but with 40 cores in parallel and with an additional threshold that the runs are stopped as soon as the first individual with a fitness less than the literature value is born. [f] Additionally, 200 individuals were created during initialization, to establish the steady-state pool.

small systems in sPuREMD, and further sPuREMD-initializations for frequent calls without the MD-simulation part to iron out the interaction between sPuREMD and OGOLEM. We note that the performance comparison between old and new code in general is highly dependent on the optimization problem, and especially on the training set. This is illustrated next in Section Disulfide application example, where we observe that this SiOH test case is not typical but, according to our experience so far, apparently provides a lower bound to the attainable speedups with OGOLEM but an upper bound with respect to algorithmic improvements of the fitness progression.

In summary, Table 1 documents that the substantially improved efficiency of our new OGOLEM/sPuREMD codebase is the combined result of (1) the algorithmic power of OGOLEM including its new extensions presented in this paper (2) the better wall-clock timings of the high-performance ReaxFF implementation sPuREMD, and (3) the linear scaling achieved by our enhanced memory management scheme (Section Scaling).

Disulfide application example

To compare the performance of the older ADF-based GA implementation with the most recent version of adaptive OGOLEM interfaced with sPuREMD in a real-life setting, a representative optimization problem was chosen from the applications currently done in the Hartke group. The molecular system contains a disulfide moiety connecting two aromatic systems, dubbed "RSSR" below. The feature of interest in a future ReaxFF study is the homolytic dissociation of the disulfide

bond, upon mechanochemical activation. The benchmark problem used features 531 molecular structures and 1765 items in the training set. These items comprise 189 atomic charges, 1089 internal coordinates, and 487 energies. The total of number of parameters to be optimized is $n_{\text{params}} = 131$.

The reference data was calculated on the RIMP2/cc-pVDZ level of theory with the ORCA program package.^[74–76] The geometries were optimized with tight convergence criteria, and the charges were calculated with the CHELPG^[77] module that employs an ESP fitting routine. Molecular structures for the energy information were taken from thermal trajectories on the semiempirical PM6 level of theory^[78,79] at different temperatures between 100 and 500 K. The PM6 trajectories were calculated using the GAUSSIAN09 suite.^[80] From these trajectories, a random set of 500 structures was taken as input for single-point calculations with the RIMP2/cc-pVDZ method. A few structures that showed convergence problems with the MP2 method were excluded from the set, therefore the total number of single point evaluations was 487.

All optimizations were run in parallel on ten cores with 40 gigabytes random access memory available. Different batches of calculations were performed with varying input parameters. Every batch contains ten identical calculations to obtain reliable averages for the optimization results. The results of these runs are compiled in Table 2. Except for *run4* and *run5* the iteration numbers were set to 20×10^3 . The *run4* and *run5* calculations were propagated for 300×10^3 iterations to get wall clock times comparable to those for the ADF runs. In all calculations, there were additional 300 evaluations to initialize the steady-state pool with random vectors. The initial force field chosen

Table 2. Comparison of absolute (wall-clock) timing results and the objective value reached with different OGOLEM input setups and with our old GA implementation; standard deviations are given in parentheses.

	Old code ^{[a],[b]}	run1 ^{[c],[b]}	run2 ^{[d],[b]}	run3 ^{[e],[b]}	run4 ^{[b],[f]}	run5 ^{[b],[g]}
∅ Timing (h)	80.5(24.1)	4.9(0.6)	5.1(0.4)	5.2(0.1)	79.3(8.4)	53.8(2.5)
Min. timing (h)	40.1	4.0	4.4	5.1	69.3	51.3
Max. timing (h)	100.7	5.5	5.6	5.3	94.2	59.5
∅ Fitness (10^4)	17.9(1.5)	19.0(1.4)	18.7(1.3)	16.3(0.7)	8.5(0.8)	7.9(1.0)
Min. fitness (10^4)	16.3	18.2	17.0	15.2	7.4	6.6
Max. fitness (10^4)	20.6	22.9	20.9	17.2	9.6	9.8

For further explanations see text. [a] GA/ADF implementation of Ref. [34]. [b] 10 cores (threads) and 40 GB memory in parallel on 4×AMD Optreron 6274 16-Core, 2.2 GHz with 32×DDR3 PC1333 Reg. ECC. [c] GA-setup closely resembles the ideal settings of the old code as found in Ref. [34]. [d] Minimal input for OGOLEM, all settings are default except for the ranges of the search space. [e] GA-setting that is currently considered ideal for OGOLEM and the problem at hand. [f] Same operator settings as *run3* but significantly bigger search space. 300,000 iteration steps were taken to get a wall time comparable to our old code. [g] Same operator settings as *run4* but with ImmediateFallback switched on.

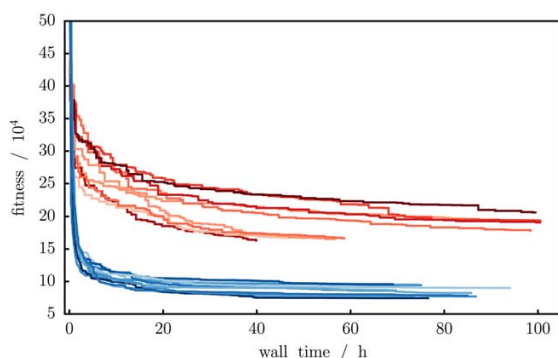


Figure 5. 10 selected but representative GA-runs with our old code (red lines)^[34,35] for the RSSR problem, in comparison to 10 with our new OGOLEM/sPuReMD-combo (blue lines, present work). The progressions of the objective function values of the ADF based runs ("old code") and the calculations of *run4* are plotted over the wall time. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

as the starting point for the optimizations was the glycine parametrization by Rahaman et al.^[81] Two sets of parameter ranges were used for the calculations. The first set features ranges of $\pm 10\%$ around the current parameter values. The second set has parameter ranges as previously used by van Duin.^[34] Both sets were corrected for inconsistencies in the parameter values. Due to the random initialization of the GA, parameter values may arise that result in single bonds being shorter than multiple bonds, which would be unphysical. Such inconsistencies were ruled out before starting the calculations.

The input settings for the ADF based calculations were those considered ideal by Larsson et al.^[34] For *run1*, the input was prepared to resemble the ADF settings as closely as possible. This comprises a single point crossover with a uniformly distributed cutting point, and a random-value multiple-parameter mutation operator. *run2* represents a minimal-input optimization with OGOLEM. The default settings chosen by the program are a single-point crossover with a Gaussian-shaped distribution of the cutting point, and a random-value multiple-parameter mutation operator. For *run3*, *run4*, and *run5*, the settings were tuned to get the best possible optimization results for the problem at hand. A mixture of 80% multipoint exchange crossover and 20% mixing recombination was used as crossover operator. The number of cuts was set to 30 ($\approx 25\%$ of n_{params}), the number of mixes was 25 ($\approx 20\%$ of n_{params}), respectively. The mutation operator was an even mixture of random-value multiple-parameter mutation and a Gaussian-weighted random-value generation around the current values of several parameters. Additionally, niching^[45] was employed. For the niching, the parameter space was divided into 20 slices per dimension. The genotypes of two individuals are defined to be in a different niche when they differ by 15 or more slices. In *run3*, 15 individuals and in *run4* 10 individuals per niche were allowed at most, respectively. This setup was found to return the best results in preliminary calculations. For *run5* the ImmediateFallback option was employed to get even better runtimes while retaining the good results of *run4*.

The result overview in Table 2 shows the overall much shorter runtimes of the new implementation for the RSSR-problem. Depending on the setup of the GA, speedups between 15.0 and 16.4 by especially taking advantage of the OGOLEM training set handling were observed (cf. lower bound in Section SiOH benchmark), therefore OGOLEM/sPuReMD can cover significantly more steps than GA/ADF within the same wall time. When the ImmediateFallback option is applied, speedups up to 22.5 were observed. Since the ImmediateFallback is invoked more often the further the calculation is propagated, even higher speedups may be obtained. However, as this acceleration is accomplished by effectively skipping unnecessary steps of the computation of the error sum, any direct graphical comparison to ADF-based runs would be meaningless and is therefore avoided altogether. Another appealing feature of ImmediateFallback is the direct elimination of items that show convergence problems in the QEq-routine or in the geometry optimizations from the pool. Therefore, parameter sets with badly misaligned parameters do not remain in the population. This leads to overall more stable results of higher quality.

The convergence behavior of the objective function value plotted over the wall time is shown in Figure 5. The superior computing time per individual results in a substantially faster convergence toward the final fitness value when using the new code. Even though the search space used in *run4* is far bigger, the fitness is almost converged after 20 h ($\approx 100 \times 10^3$ iterations). At the same time, the ADF-based GA shows no signs of convergence at all. Furthermore, no ADF-based calculation could be completed within the large search space. GA/ADF runs into trouble for the error-sum evaluation for most individuals with this setting, which ultimately leads to premature termination of the run. If the final fitness value is taken into consideration, another superiority of the new code becomes apparent. The crossover and mutation operators implemented in OGOLEM give even better optimization results than the already well-performing GA/ADF code.^[34] The final fitness values of each optimization are shown in Table 2. The final results of *run1* and *run2* are worse than in the GA/ADF reference runs. In case of *run1* the difference may be explained by differences between the OGOLEM code and GA/ADF. The user-defined input is only part of the GA-parameters that determine the general performance of the algorithm, and the results react quite sensitively to the setup of the GA. Therefore, the settings for the optimization are not completely interchangeable between the various implementations. The default setting used in *run2* employs single-point crossover with a Gaussian-shaped distribution of the cutting point, which is not very well suited for the present optimization problem.^[19] It was thus expected that the final force fields would be inferior to the GA/ADF results. Nevertheless, the results are reasonably close to the best ones obtained and therefore would be a fine starting point for users lacking experience with GA. Since the results of the optimization rely heavily on the input, as argued above, the settings for *run3* and *run4* were chosen more carefully. Using the new mixing recombination operator and niching it was possible to obtain even better performance per

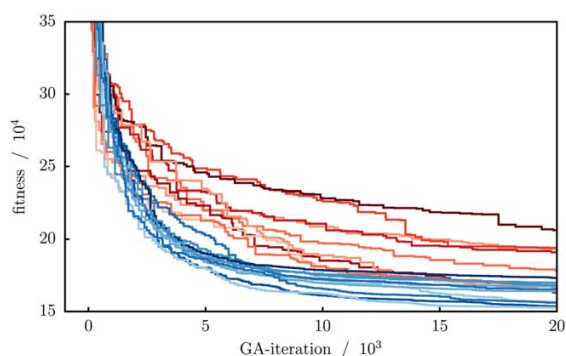


Figure 6. 10 selected but representative GA-runs with our old code (red lines)^[34,35] for the RSSR problem, in comparison to 10 with our new OGOLEM/sPuReMD-combo (blue lines, present work). The progressions of the objective function values of the ADF based runs (“old code”) and calculations of *run3* are plotted over the number of GA-iterations.

iteration than with our old ADF-based code. Figure 6 shows the comparison of the objective function value plotted vs. the iteration number for both codes with ideal settings.

Thus, the favorable features of the new code compared with the old one can be traced back to (1) the substantial leap in computing performance, that is, for this training set even up to a factor of 16 (or 22.5 with `ImmediateFallback` and still below the regime of additional scaling benefits that are reached for more than 10 threads), (2) the new algorithmic details explained above (the operator settings do have an impact but default GA-settings that are qualitatively different to the older code already bring in much of the overall improvement), and (3) the better usability and stability. Therefore, OGOLEM/sPuReMD solves a lot of problems associated with limitations of computing resources and shifts the focus of the user more towards the quality of the reference data and the choice of the ReaxFF parameter set. In fact, for these RSSR systems, ongoing work in our labs is devoted to improving strat-

egies for training set creation and validation, as well as to molecular dynamics simulations of these mechanoswitchable system. Results for that will be reported in future publications.

Scaling

As an illustration of linear scaling we achieve with our OGOLEM/sPuReMD-combo, Figure 7 shows strong scaling results of the SiOH and RSSR problems discussed above. To avoid distracting scatter and artifacts from our intrinsically nondeterministic algorithms, these scaling tests were artificially restricted to no parameter variations at all. Thus, in these tests, no minimization of the objective function happens, but nevertheless all calculational steps are performed exactly as in a production mode. Additionally, besides the artificial “deterministic” zero-dimensional search space, all other settings (population size, GA iteration number, GA operators, etc.) also correspond to choices that would be made for production. Therefore, Figure 7 displays the true scaling underlying actual real-life GA calculations. The figure shows acceleration factors as a function of used threads (equal to the number of used CPU cores) for up to 48 threads, and normalized to the timings of the single-thread runs. Only the true global optimization part is taken into account; the initial start-up and pool-filling stages are not included.

Figure 7 clearly illustrates that the parallelization at the GA level in OGOLEM leads to linear scaling in practice (red curves), with sPuReMD as ReaxFF backend. These scaling characteristics have already been observed in previous OGOLEM applications to different optimization tasks, for example with cluster structure optimization,^[40,59] parameter fitting with traditional force fields^[59] and abstract benchmarks.^[62] Therefore, it can be taken as an intrinsic feature of the OGOLEM architecture.

Nevertheless, care has to be taken to not destroy this feature with new backends: The linear scaling shown in Figure 7 pertains only when the newly implemented memory management with scratch spaces for the sPuReMD backend (discussed

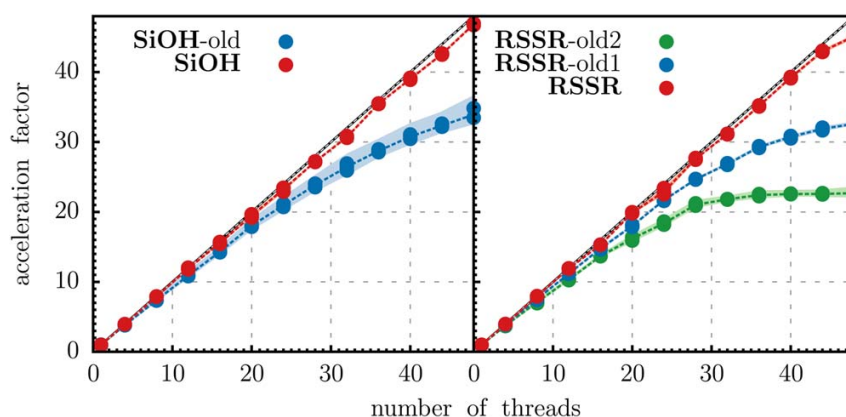


Figure 7. Strong-scaling benchmarks of the SiOH optimization problem (SiOH, compare Section SiOH benchmark) and a training set for the RSSR problem (similar to that of Section Disulfide application example) using shared-memory parallelism. Versions with “old” in their names are shown for comparison, they are not the results of the final implementation (see text). All calculations were performed three times each with the given number of threads/cores (1, 4, 8, ...). Small fluctuations of acceleration are illustrated by the spread in the light color lines; the size of this spread is similar to the size of the dots on the line.

above in Section The *ogolem-sPuReMD* combination) is actually used. Without this improvement, severe memory allocation bottlenecks thwart the potentially linear scaling already at moderate numbers of threads (10–30, green and blue curves). Moreover, this is problem dependent, as illustrated by the training sets for **RSSR-old1** and **RSSR-old2** that are different, that is, **RSSR-old1** is the same as **RSSR**, but the calculation of the former did not use that memory management. The same holds for **SiOH** and **SiOH-old**. **RSSR-old2**, however, is mainly a smaller training set with less items and just many single points leading to many backend-calls in a smaller amount of time and without the new memory management, too. Nevertheless, also this **RSSR-old2** setting can be calculated within linear scaling with the new memory management of *sPuReMD* (not shown). Thus, all investigated optimizations via *OGOLEM/sPuReMD* (many more than shown here) have this scaling behavior without problem dependence now.

Using shared memory, this linear acceleration for our training set calculations was not possible with the old GA/ADF code that employed MPI parallelization at the reference-item. The old setup was hampered by several problems, including (1) hardly avoidable overload of the master process handing out calculation tasks to the slaves, due to huge time differences of these tasks, and (2) additional locks and serial bottlenecks since different reference item calculations depended on each other. Thus, as remarked in Ref. [34], the scaling of our old GA/ADF setup was good for small numbers of cores but became inefficient rather quickly (between 16 and 32 cores). In contrast, our new *OGOLEM/sPuReMD-combo* can still be used efficiently with significantly higher numbers of cores. Therefore, parallelism can be conveniently used to combat both lack of domain-specific *a priori* knowledge and search space difficulty (cf. Section Objective function surface).

Related Work

There has been some previous work in the literature on GA-use for ReaxFF parameter optimization. In this section, we briefly discuss the relations between the prior work and our present contribution.

Parameters in a specialized charge-transfer force field^[82] were optimized with a GA.^[31,32] Pahari and Chaturvedi^[33] optimized ReaxFF parameters with a GA, but the focus of their paper was on determining a minimal set of parameters to vary in the GA based on prior sensitivity tests and cross-correlations.

Jaramillo-Botero et al. have also used a GA to optimize ReaxFF parameters^[36]; however, they did not use crossover steps, only mutation, had limited possibilities for parallelization, and only aimed at adding 37 parameters for a chlorine atom to an already established ReaxFF for Si-, C-, and H-atoms. In contrast, in Ref. [34], between 67 and 191 parameters for three atoms (Si, O, H) were varied simultaneously, using a full-blown GA with parallelization across reference data items. Additionally, we have applied the same program suite to the photochemical isomerization of azobenzene,^[35] generating a purely force-field-based model for nonadiabatic transitions between two electronic states and simultaneously exploring

the real-life case of ReaxFF parameter optimization with little prior knowledge about needed reference data items and suitable parameter ranges.

Shortly before the present article was submitted, a pair of papers by Weingarten et al.^[37,38] was published. These authors reoptimized 46 parameters of a previously published ReaxFF parametrization for two explosives, using mutation-only evolutionary strategies in a multi-objective setting. The latter is advertised as getting rid of the necessity to attach a predefined weight to each training set item. However, for the about 3600 items in their training set, they actually retained most of the predefined weights; only five values (relative weights between different, large item-groups) were left open. We suspect that this is necessary to keep population size and search space dimensionality practically manageable, despite the use of supercomputers. Nevertheless, post-selection of suitable candidates from the five-dimensional Pareto front apparently became an issue. For these reasons, we believe that single-objective EA approaches (as used here) will remain competitive.


Conclusions

By joining *OGOLEM* and *sPuReMD*, two advanced implementations of GA and of the reactive force field ReaxFF, respectively, we have significantly improved upon the efficiency and usability of reactive force field generation. Particular care was taken to retain the theoretically excellent scalability of GAs, to enable future massively parallel usage of this code combination. For both benchmark and real-life examples, we have demonstrated clear superiority of our present implementation over our earlier one,^[34] despite the successes of the latter.^[34,35] This progress directly translates into advantages for the end user, as it brings needed real times for typical ReaxFF global parameter optimization tasks from weeks down to hours, and from multiple cascading runs with in-between adjustments by the user down to single runs of black-box character.

We are confident that these improvements in global force-field parameter optimization will also make future research on how to choose training sets and how to validate force-field performance easier.

Keywords: reactive force fields · ReaxFF · global optimization · genetic algorithms

How to cite this article: M. Dittner, J. Müller, H. M. Aktulga, B. Hartke. *J. Comput. Chem.* **2015**, *36*, 1550–1561. DOI: 10.1002/jcc.23966

 Additional Supporting Information may be found in the online version of this article.

- [1] D. Marx, In *Computational Nanoscience: Do It Yourself!* J. Grotendorst, S. Blügel, D. Marx, Eds.; John von Neumann Institute for Computing: Jülich, **2006**; p. 195.
- [2] J. Hutter, *WIREs Comput. Mol. Sci.* **2012**, *2*, 604.

- [3] S. Piana, K. Lindorff-Larsen, D. E. Shaw, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5915.
- [4] A. Nakano, R. K. Kalia, K.-I. Nomura, A. Sharma, P. Vashishta, F. Shimojo, A. C. T. van Duin, W. A. Goddard, R. Biswas, D. Srivastava, L. H. Yang, *Int. J. High Perf. Comput. Appl.* **2008**, *22*, 113.
- [5] A. Warshel, *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425.
- [6] W. Thiel, H. M. Senn, *Angew. Chem. Int. Ed.* **2009**, *48*, 1198.
- [7] R. Mata, *Phys. Chem. Chem. Phys.* **2010**, *12*, 5041.
- [8] K. Farah, F. Müller-Plathe, M. C. Böhm, *Chem. Phys. Chem.* **2012**, *13*, 1127.
- [9] D. W. Brenner, *Phys. Rev. B* **1990**, *42*, 9458.
- [10] J. Hur, S. J. Stuart, *J. Chem. Phys.* **2012**, *137*, 054102.
- [11] B. C. Bolding, H. C. Andersen, *Phys. Rev. B* **1990**, *41*, 10568.
- [12] J. Aqvist, A. Warshel, *Chem. Rev.* **1993**, *93*, 2523.
- [13] F. Jensen, P.-O. Norrby, *Theor. Chem. Acc.* **2003**, *109*, 109.
- [14] J. Danielsson, M. Meuwly, *J. Chem. Theory Comput.* **2008**, *4*, 1083.
- [15] K. D. Smith, S. I. Stoliarov, M. R. Nyden, P. R. Westmoreland, *Molec. Simul.* **2007**, *33*, 361.
- [16] T. Liang, Y. K. Shin, Y.-T. Cheng, D. E. Yilmaz, K. G. Vishnu, O. Veners, C. Y. Zou, S. R. Phillpot, S. B. Sinnott, A. C. T. van Duin, *Annu. Rev. Mater. Sci.* **2013**, *43*, 409.
- [17] J. A. Martinez, D. E. Yilmaz, T. Liang, S. B. Sinnott, S. R. Phillpot, *Curr. Opin. Solid State Mater. Sci.* **2013**, *17*, 263.
- [18] F. Avaltroni, C. Corinboeuf, *J. Comput. Chem.* **2011**, *32*, 1869.
- [19] T. Weise, Global Optimization Algorithms—Theory and Application, Available at: <http://www.it-weise.de/projects/>, **2011**, Last accessed 9 June 2015.
- [20] J. Hunger, S. Beyreuther, G. Huttner, K. Allinger, U. Radelof, L. Zsolnai, *Eur. J. Inorg. Chem.* **1998**, *6*, 693.
- [21] J. Hunger, G. Huttner, *J. Comput. Chem.* **1999**, *20*, 455.
- [22] T. R. Cundari, W. T. Wi, *Inorg. Chim. Acta* **2000**, *300*, 113.
- [23] B. Courcot, A. J. Bridgeman, *J. Comput. Chem.* **2011**, *32*, 240.
- [24] T. Strassner, M. Busold, W. A. Herrmann, *J. Comput. Chem.* **2002**, *23*, 282.
- [25] M. Tafipolsky, R. Schmid, *J. Phys. Chem. B* **2009**, *113*, 1341.
- [26] J. M. Wang, P. A. Kollman, *J. Comput. Chem.* **2001**, *22*, 1219.
- [27] A. Globus, M. Menon, D. Srivastava, *Comput. Model. Eng. Sci.* **2002**, *3*, 557.
- [28] C. R. Herbers, K. Johnston, N. F. A. van der Vegt, *Phys. Chem. Chem. Phys.* **2011**, *13*, 10577.
- [29] B. C. Barnes, L. D. Gelb, *J. Chem. Theory Comput.* **2007**, *3*, 1749.
- [30] C. M. Handley, R. J. Deeth, *J. Chem. Theory Comput.* **2012**, *8*, 194.
- [31] L. Angibaud, L. Briquet, P. Philipp, T. Wirtz, J. Kieffer, *Nucl. Instrum. Meth. B* **2011**, *269*, 1559.
- [32] L. G. V. Briquet, A. Jana, L. Mether, K. Nordlund, G. Henrion, P. Philipp, T. Wirtz, *J. Phys.-Condens. Mat.* **2012**, *24*, 395004.
- [33] P. Pahari, S. Chaturvedi, *J. Mol. Model.* **2012**, *18*, 1049.
- [34] H. R. Larsson, A. C. T. van Duin, B. Hartke, *J. Comput. Chem.* **2013**, *34*, 2178.
- [35] Y. Li, B. Hartke, *J. Chem. Phys.* **2013**, *139*, 224303.
- [36] A. Jaramillo-Botero, S. Naserifar, W. A. Goddard, III, *J. Chem. Theory Comput.* **2014**, *10*, 1426.
- [37] J. P. Larentzos, B. M. Rice, E. F. C. Byrd, N. S. Weingarten, J. V. Lill, *J. Chem. Theory Comput.* **11**, 2015, 381
- [38] B. M. Rice, J. P. Larentzos, E. F. C. Byrd, N. S. Weingarten, *J. Chem. Theory Comput.* **11**, 2015, 392.
- [39] H. M. Aktulga, S. A. Pandit, A. C. van Duin, A. Y. Grama, *SIAM J. Sci. Comput.* **2012**, *34*, 1.
- [40] J. M. Dieterich, B. Hartke, *Mol. Phys.* **2010**, *108*, 279.
- [41] J. M. Dieterich, B. Hartke, Available at: <http://www.ogolem.org/>, Last accessed 9 June 2015
- [42] B. Hartke, *Angew. Chem. Int. Ed.* **2002**, *41*, 1468.
- [43] B. Hartke, *WIREs Comput. Mol. Sci.* **2011**, *1*, 879.
- [44] B. Hartke, *J. Phys. Chem.* **1993**, *97*, 9973.
- [45] B. Hartke, *J. Comput. Chem.* **1999**, *20*, 1752.
- [46] B. Hartke, *Z. Phys. Chem.* **2000**, *214*, 1251.
- [47] B. Hartke, H.-J. Flad, M. Dolg, *Phys. Chem. Chem. Phys.* **2001**, *3*, 5121.
- [48] F. Schulz, B. Hartke, *Chem. Phys. Chem.* **2002**, *3*, 98.
- [49] B. Hartke, *Phys. Chem. Chem. Phys.* **2003**, *5*, 275.
- [50] A. Tekin, B. Hartke, *J. Theor. Comput. Chem.* **2005**, *4*, 1119.
- [51] B. Bandow, B. Hartke, *J. Phys. Chem. A* **2006**, *110*, 5809.
- [52] B. Hartke, *Chem. Phys.* **2008**, *346*, 286.
- [53] J. M. Dieterich, U. Gerstel, J.-M. Schröder, B. Hartke, *J. Mol. Mod.* **2011**, *17*, 3195.
- [54] U. Buck, C. C. Pradzynski, T. Zeuch, J. M. Dieterich, B. Hartke, *Phys. Chem. Chem. Phys.* **2014**, *16*, 6859.
- [55] F. Koskowski, B. Hartke, *J. Comput. Chem.* **2005**, *26*, 1169.
- [56] B. Hartke, *Chem. Phys. Lett.* **1996**, *258*, 144.
- [57] B. Hartke, *Theor. Chem. Acc.* **1998**, *99*, 241.
- [58] B. Hartke, M. Schütz, H.-J. Werner, *Chem. Phys.* **1998**, *239*, 561.
- [59] J. M. Dieterich, B. Hartke, *J. Comput. Chem.* **2011**, *32*, 1377.
- [60] H. R. Larsson, B. Hartke, *Comput. Meth. Mater. Sci.* **2013**, *13*, 120.
- [61] N. O. Carstensen, J. M. Dieterich, B. Hartke, *Phys. Chem. Chem. Phys.* **2011**, *13*, 2903.
- [62] J. M. Dieterich, B. Hartke, *Appl. Math.* **2012**, *3*, 1552.
- [63] A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, III, *J. Phys. Chem. A* **2001**, *105*, 9396.
- [64] A. C. T. van Duin, J. M. A. Baas, B. van de Graaf, *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 2881.
- [65] A. D. Kulkarni, D. G. Truhlar, S. G. Srinivasan, A. C. T. van Duin, P. Norman, T. E. Schwartzentruber, *J. Phys. Chem. C* **2013**, *117*, 258.
- [66] A. K. Rappe, W. A. Goddard, *J. Phys. Chem.* **1991**, *95*, 3358.
- [67] H. M. Aktulga, J. C. Fogarty, S. A. Pandit, A. Y. Grama, *Parallel Comput.* **2011**, *38*, 245.
- [68] J. C. Fogarty, H. M. Aktulga, A. Y. Grama, A. C. T. van Duin, S. A. Pandit, *J. Chem. Phys.* **2010**, *132*, 174704.
- [69] M. Yusupov, E. C. Neyts, C. C. Verlackt, U. Khalilov, A. C. T. van Duin, A. Bogaerts, *Plasma Process. Polym.* Doi: 10.1002/ppap.201400064.
- [70] S. B. Kylasa, H. M. Aktulga, A. Y. Grama, *J. Comput. Phys.* **2014**, *272*, 343.
- [71] A. Y. Grama, H. M. Aktulga, S. B. Kylasa, Available at: www.cs.purdue.edu/puremd, Last accessed 9 June 2015.
- [72] S. Liang, The Java Native Interface: Programmer's Guide and Specification, Addison-Wesley, Reading, Massachusetts, **1999**.
- [73] A. C. T. van Duin, A. Strachan, S. Stewman, Q. Zhang, X. Xu, W. A. Goddard, III, *J. Phys. Chem. A* **2003**, *107*, 3803.
- [74] F. Neese, *WIREs Comput. Mol. Sci.* **2012**, *2*, 73.
- [75] F. Neese, *J. Comput. Chem.* **2003**, *24*, 1740.
- [76] S. Kossmann, F. Neese, *J. Chem. Theory Comput.* **2010**, *6*, 2325.
- [77] C. M. Breneman, K. B. Wiberg, *J. Comput. Chem.* **1990**, *11*, 361.
- [78] J. J. P. Stewart, *J. Mol. Model.* **2007**, *13*, 1173.
- [79] J. J. P. Stewart, *J. Mol. Model.* **2009**, *15*, 765.
- [80] Gaussian 09, Revision D.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian, Inc., Wallingford CT, **2009**.
- [81] O. Rahaman, A. C. T. van Duin, W. A. Goddard, III, D. J. Doren, *J. Phys. Chem. B* **2010**, *115*, 249.
- [82] L. Huang, J. Kieffer, *J. Chem. Phys.* **2003**, *118*, 1487.

Received: 12 March 2015
Revised: 9 May 2015
Accepted: 15 May 2015
Published online on 17 June 2015

4.3 Complementary Information about Parallelization Improvements

In computer simulations in the context of natural sciences, clearly the studied physics (or chemistry) is the *essential* part. Notwithstanding, especially when working with bigger program packages, more elaborated programming skills and techniques are usually required (or *should* be accomplished), which directly places the actual daily work into the regime of computer science and software engineering. As this Thesis shall not be a general introduction into any such matters, only *one* typical example of a design decision/change is given in the following.

Indeed, a big part of all the work within this Thesis is programming: Starting already with a framework as OGOLEM, which has been highly extended since its foundation by DIETERICH (*cf.* Chapter 3) and which provides already very many features *somewhere* in the program, the main challenge is to combine all those features in a way that leads to the successful global optimization of the new chemical tasks, besides implementing new ones. The highest priority then is to follow the frameworks' given structure or "logic" without "reinventing the wheel", but incorporating some generifications^[434] and generalizations that are meaningful. Also the—in the author's opinion—high standard of code quality, following a well-thought **object orientation (OO)** design and other programming techniques/principles (*cf.* the depictions of DIETERICH^[389]), should be approached in any further extension. In this regard, the present complementary information shall shed light on one (small) implementation that was needed to improve the parallelization of the REAXFF parameter optimization task.

The main implementations of REAXFF optimization took place during the Master's Thesis of the present author.^[244] However, back then, the **SMP** parallelism known to be (almost) linearly scaling in usual **GA** applications or production modes^[389,390,404,405] was unexpectedly worse. The needed improvements followed then at the beginning of the work for this Thesis.

The central parallelism in OGOLEM which is based on the Java threads concurrency system¹ is working on the level of **GA** cycles as already described in Section 3.3. By this structure based on the pool-**GA**,^[422] the "trivial parallelism" of **GA** optimization follows. Having a look at AMDAHL's Law^[436] as a pessimistic assessment of maximal (theoretical) speedup, this simple limiting case scenario leads to

$$\text{speedup}(r_p, N) = \frac{1}{1 - r_p + \frac{r_p}{N}}, \quad (4.1)$$

with N as the number of parallel executions and r_p as the ratio of parallel portion of the program; $r_s = 1 - r_p$ is the sequential (serial) part then.² Besides the pessimistic view of not

¹ Because of the platform-independency and portability of Java, such multithread-parallelism is integrated directly into the language and in abstract form determined in the official language specifications.^[435] How this maps to the thread-system of the operating system (in usual Unix-derivatives and the other systems) is then defined by the used Java implementation and carried out by the **Java virtual machine**.

² Note the well-known restrictions of this view, as by GUSTAFSON,^[437] usually the problem size changes with

included effects, this sets the scene for two further corollaries: If the parallel fraction, r_p , is small, an increase of N will make no sense. The problem execution time is completely dominated by the serial part then. If N approaches infinity, the maximal speedup is bound by $1/(1-r_p)$ from Eq. (4.1), or looking at the execution time themselves: If the overall execution time were, e.g., $T = 20$ h with $N = 1$ in $T = Tr_s + T\frac{r_p}{N}$, with $r_s = 5\%$ serial ratio, it would become and *stay* 1 h at $\lim_{N \rightarrow \infty}$, but no less. Thus, the scalability of the program is driven by the ratio of the program parts that must be executed serially.^[440] Assuming an unreal $r_p = 1$ for the sake of the argument, Eq. (4.1) yields this linear scaling as the speedup(N) = N .

As long as the *serial* portion in OGOLEM which extends (almost exclusively) to the part of pool-additions is minuscule compared to the rest, this behavior of linear scaling is to be expected. The *serial* part of OGOLEM *during the main GA optimization* is the task to decide whether to add a created new candidate solution to the pool—the shared data—and therefore must be locked against (or synchronized with) further concurrent additions/changes by other threads at the same time. On the one hand, these serial checks usually only include some fast fitness and niching comparisons, *i.e.*, mainly the result handling after a GA cycle, besides some internal pool's state updates.³ The disparate parallel part, on the other hand, includes all other operations, and most importantly, the fitness evaluation; all these separate operations can come with, e.g., many steps, such as MD snippets, local optimizations of chemical structures or of the fitness function itself, etc. Thus, all the computations in the parallel part are completely dominating in any real application settings, usually.⁴

Since the *relation* between *serial* and *parallel* executions in OGOLEM were not changed by implementing the sPUREMD interface and fitness calculations, the observed break down of the linear scaling could instead be traced back to the *memory management*.⁵ Here, each call of sPUREMD leads to small dynamic allocations of memory in the C part for all internal data structures needed for REAXFF computations. Although these can be quite small (about tens to hundreds of KBs, very dependent on the chemical systems), this was already quite too much when multiple separate threads try to dynamically (de-)allocate such memory for *each single call* of sPUREMD, for a call that needs time (for small systems) also *just* on the order of microseconds.⁶ This underpins the fact that a very loose combination without further

regard to the available parallelism (processors), *i.e.*, r_p is dependent on N . Tackling bigger problems can therefore dependently increase the parallelizable part. Also, other overhead due to communication between different executions and scheduling is not included here. Further perspectives of, e.g., multicore architectures and power management in the context on AMDAHL's law can be found in Refs. [438, 439].

³ Certainly, also the shared resources by the concurrent queuing of the separate tasks introduces shared data and some serial part.

⁴ Just as limiting case *gedankenexperiment*: Of course, if one used a ridiculously giant pool and/or used/implemented very expensive niche comparisons by using the non-fixed “dynamic” niching (*cf.* Chapter 5 and Fig. 3.7) and at the same time very cheap GA operators with a very tiny (one SP) training set for REAXFF, this relation would not hold anymore and the parallelization would break down. As a side note, this is also the reason for not having bothered yet to use something as HC itself for niching, as it was proposed in Ref. [365], because the clustering itself would take place in the *serial* part.

⁵ For the memory management, the operating system usually supplies an allocator for the C library functions `malloc()/calloc()`, which were also changed to known other ones, as for instance, `jmalloc`,^[441] since these were invented for the reason of better scalability in multithreading (and less memory fragmentation).^[240] Those changes, however, did also not solve the additional allocation bottleneck that was apparently introduced by too frequent and (accumulated) huge (de-)allocations.

⁶ Though, with all the other computations in the parallel part, including extended training sets, the parallel GA

changes of sPuDREMD and OGOLEM by a small **Java native interface (JNI)**^[442] layer—the communication interface between Java and C—for energy/gradient calls on REAXFF level of theory can thwart the linear scaling, and therefore, a more extensive implementation was meaningful. At the end, this both needed very low-end changes to sPuDREMD and high-end ones to OGOLEM.⁷

General caveat for interfacing: This relation between serial vs. parallel can always become maladjusted when using external (usually stand-alone) programs with very short overall executions times. Also in other contexts, as for instance, **SQC** calculations for *one SP* of a small chemical system as in **GOCAT** design, just the plain *start-up* time of the external program, including any (de-)allocations of memory and often some I/O, can also worsen the parallel performance. However, in such cases, one simply does not always want to—or is able to, without source code—migrate to an **application programming interface (API)** or even create one that allows a start-up of the program once, for each thread, and that allows subsequent **SP** computations without further unnecessary overhead. If possible, the work in the external program should be enlarged, *e.g.*, not by just calculating *one* fast **SP**, but maybe by using one external *local optimization* or calculating *multiple separate* chemical systems at once. When using internal energy/gradient backends in OGOLEM, which provides some empirical potentials (see Ref. [413]), caching/scratch-space without further redundant overhead is always used, of course.

Final strategy of implementation: In the case of interfacing sPuDREMD, the final strategy of parallel execution improvements exactly evolved to implementing something as: (1) a start-up of the external REAXFF interface sPuDREMD, (2) then an idle mode waiting for computations without any further overhead with completely allocated work or scratch space. Because of the division in *native* memory in java and the one the **Java virtual machine (JVM)** supervises, some more care must be taken here; usually all C memory must be handled in the native part as the **JVM** does not manage that. This was used as intended “memory leak” from the perspective of the Java part, *i.e.*, explicitly unfreed memory, but that was remembered by explicit manual pointer (arithmetic) based address communication between C and Java.

In the following, more details and thoughts are presented on how this was achieved. The comparison of the broken scalability *before* and the improvements *after* are shown in the publication on p. 106. The following section premises terminology and knowledge of **OO** programming in general and Java in particular.^[443,444] The descriptions are oriented along *design patterns*^[445–447] as general reusable solutions to commonly occurring problems in software design. However, design patterns are used here as common vocabulary for communication; the actual design did *not* just try to enforce patterns as an end in itself. Without knowing such vocabulary, the meaning of the steps can certainly also be grasped from the context.

step execution time usually is on the order of seconds and above and thus not deteriorating the (serial/parallel) relation.

⁷ “High-end” and “low-end” shall simply denote the abstraction level of implementation.

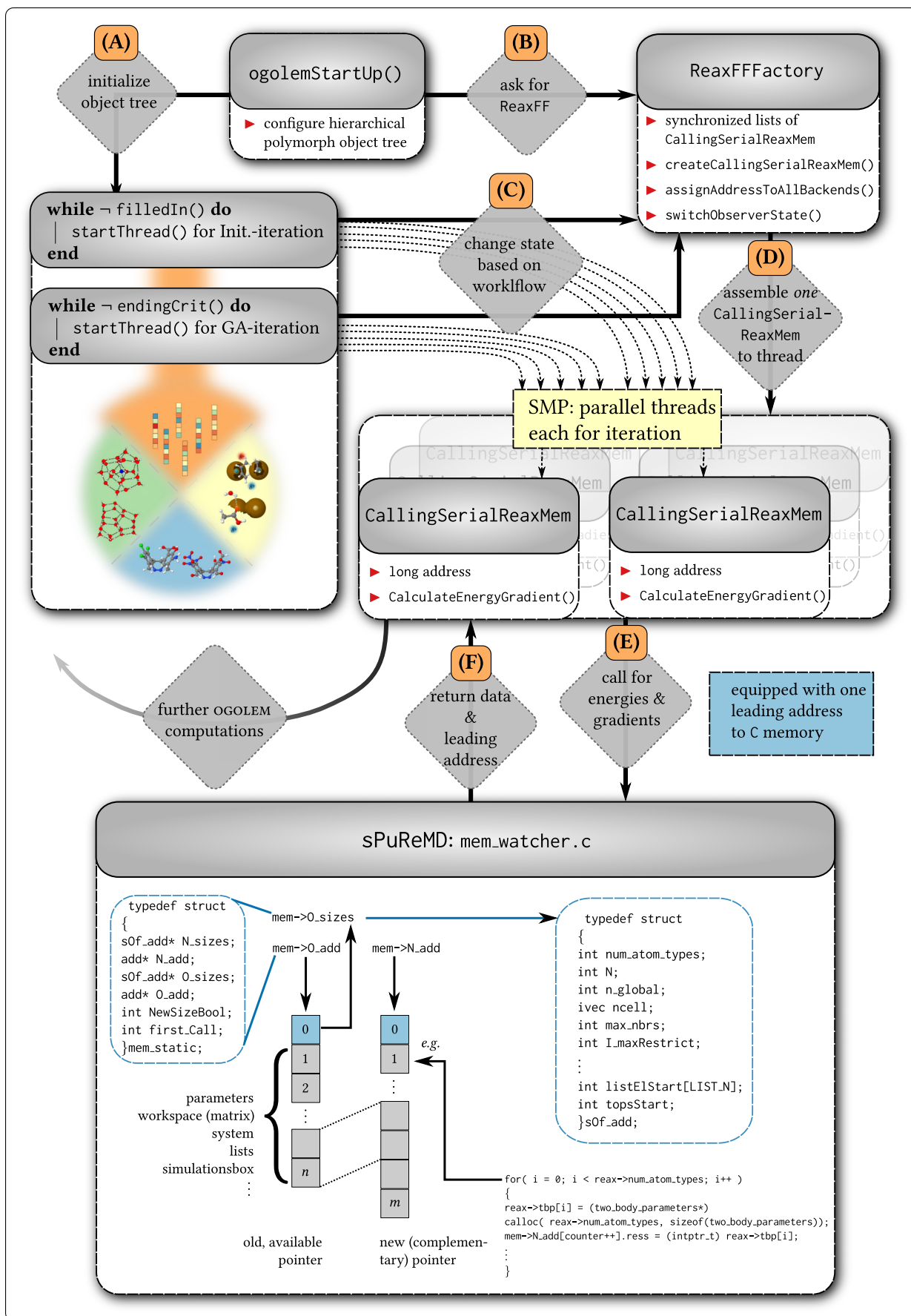


Fig. 4.2: Simplified overview of the important classes and their interactions implemented to obtain an explicit memory handling between OGOLEM and sPuReMD. See the main text for explanations.

The most important points are the following, which are illustrated in Fig. 4.2 on the previous page:

- Definition of a *utility* class, ReaxFFFactory, which exposes a Factory Method for the construction of objects of the class for calling REAXFF (see (B)), CallingSerialReaxMem), and furthermore, the management of the needed memory addresses ((D)). This class itself wraps *synchronized* lists of already constructed CallingSerialReaxMem as a Singleton.
- This relation is similar to that of an Observer pattern, as all instances of this CallingSerialReaxMem class that are instantiated are cached in the lists; this uses WeakReferences in the background in order to allow the garbage collector of Java to do its work if such a caller is not used anymore in the subsequent workflow, not to introduce a subtle memory leak here in the Java part. This means that, as long as any fitness function needs to call REAXFF, the backend is needed and its instance existence is remembered. However, an usual Observer pattern would also allow for state-based updates (notifications) from the subject, ReaxFFFactory, to its observers, which is not needed here. Instead, utility functions are exposed to manage and distribute available pointer addresses in Java to the separate threads: During the parallel execution of the initialization and/or GA tasks, which is semaphore-based and uses redundantly available scratch space object trees, such addresses are delegated to the threads again to serve as *one* address space that is used in each sPUREMD call per thread ((D)).⁸
- The utility class implements a State/Strategy that follows the general OGOLEM workflow: There is always an initialization phase of generating the starting pool. Either candidate solutions are randomly generated or older solutions, as seeds, are read in. Based on the needs of the workflow, this switched state effectively controls whether the observer is monitoring each ever instantiated CallingSerialReaxMem (which needs that additional small *serial* bottleneck due to synchronization of the lists), or whether it is not needed anymore, which is true *after* the initialization for the main GA loop ((C)).
- As important background information: OGOLEM uses deep hierarchical object structures/trees consisting of, e.g., optimizers, their backends to encode the data (coordinates/parameters/charges, etc.), fitness functions, low-end backends for energy/gradient calculations, besides other trees for GA operators, etc. These are configured at start-up time during OGOLEM *via* common Abstract Factories that compose the complete tree for further polymorphic execution. This is named ogolemStartUp() in Fig. 4.2, (A). The tree is usually copied using a construction similar to a Prototype

⁸ Some further complications are introduced, as additionally so-called ReaxFFC objects are instantiated (not shown) that use CallingSerialReaxMem for energy/gradient calculations. The former wrap a huge context-specific state for managing all types of different fitness ingredients (geometry/energy/gradient/RMSG/cell parameters, etc.). Thus, a management from many (ReaxFFC) instances to one (CallingSerialReaxMem) is needed; none of the address spaces must get lost as otherwise a memory leak of never re-used and never deallocated C memory would be present.

pattern⁹ for each thread that should run in parallel. In this way, also each such a trunk will use thread-local scratch space (and all object's state) and this will never lead to any concurrency issues of threads wanting to use the same address space or any other visibility issues of polling from main memory etc.^[440] In other words: Each parallel execution of one GA cycle (and also one initialization cycle) uses a complete, usually deep copy of all objects it needs.

- This redundant object tree, which was already used before these improvements as cache of one thread, now can be capitalized as point of entry to also distributing explicitly all the `CallingSerialReaxMem` to each execution thread. Thus, every parallel thread will only be assembled to know one `CallingSerialReaxMem` and use that for each REAXFF call ((D)); the reason for this will become apparent in a moment.
- In the C code of `sPurEMD`, each allocation and deallocation on the heap ever occurring in the program was also capsuled into one file, *i.e.*, `mem_watcher.c`. All important size information of all C structs that are allocated is separately remembered, *e.g.*, the size information named `O_sizes` and `N_sizes` in Fig. 4.2, standing for “old” and optionally “new”. All allocated pointer addresses are saved in a list, *i.e.*, `O_add/N_add` in Fig. 4.2, by explicit casts to `intptr_t`, which is used to store the pointer value platform-independently in a non-pointer type, *i.e.*, an integer that is at least as big as the largest pointer. The first entry of that list is the size information of everything that is needed (`O_sized`) and is/was allocated in C. Thus, if the first call of `sPurEMD` in one `CallingSerialReaxMem` happens ((E)), *no* memory ever was allocated before and `sPurEMD` simply allocates enough for all data structures, such as the parameters themselves, the coordinates, etc. Every pointer of each (partially multi-dimensional) structure is saved in the C-address list (`O_add`). When returning to `OGOLEM` again, it also delivers back the leading pointer address as a Java long type pointing to this address list structure ((F)). `OGOLEM` then saves the address as internal state of the wrapping `CallingSerialReaxMem`.
- In each subsequent call in that same one `CallingSerialReaxMem`, `sPurEMD` gets handed over the leading address to the address list from Java ((E)) and decides whether the allocated structures are still big enough because the size of the chemical systems per call can vary. If the size is sufficient, it will simply re-use the same memory already allocated in any antecedent call. Otherwise, memory is re-allocated by resizing the structures, and the complemented addressed are saved in that same list; this leads to the doubling in `O_add` vs. `N_add`. After some calls, each memory address list will be grown to the needed maximum size, will hence be able to treat all the systems in the training set and will thus be “configured” or fully initialized, *i.e.*, it will not change anymore. This will be true in the first part of the initialization of `OGOLEM`, already.

⁹ That is, each optimizable target such as a cluster or a parameter vector, etc., is used strictly as Prototype for cloning it and changing it slightly in subsequent GA cycles. The complete (deep) copy of object trunks cannot strictly be subsumed under the Prototype pattern if just used to merely double the object tree.

- In all subsequent calls ever occurring in OGOLEM, no C memory must be adapted anymore and the scratch space can simply be re-used: Instead of explicit `malloc/calloc` commands in C, simply reading out old addresses as cast from pointers (from `intptr_t`) to the struct actually used. This is the “idle” mode mentioned above.
- The exact recursive loops of allocations and deallocations in the C part must be perfectly matched between each call to re-use each older explicit pointer in the exact same order of those loops, without implementing further more complex execution order managements. This would obviously be very cumbersome in more complex workflows in C, but is possible here, as the sPvREMD fulfills just *one* clear purpose, *i.e.*, being an energy/gradient provider, whereas additional features such as MD simulations, etc., are switched off. Therefore, this strategy was tackled because of the well-arranged memory handling in sPvREMD, besides its other general performance improvements^[410,411] bringing in almost no further redundancies of not needed overhead.

Thus, the advantages of a good OO design can be summarized as follows: Using a Factory Method, no sidestepping the object tree by using any other way to call REAXFF can (usually) happen.¹⁰ Each class instantiation that is able to make a REAXFF call is known at all times if they are still needed and not garbage collected; these are the observers of the Observer pattern. By using the static synchronized leading address lists in Java, the lists define a Singleton at all time, which does not allow to create multiple such lists at the same time in OGOLEM, which would naturally be wrong. The subject of the observers knows and re-distributes all leading pointer addresses as known object pool to the observers and threads. These threads use all their own thread-local space in Java as well as in C now, by being equipped with one leading address. This leading address gets extracted to the size information and all other pointers in C and can simply be re-used. With the State pattern in ReaxFFFFactory, the synchronization or monitoring level can be changed based on OGOLEM’s general workflow, such that the additional *serial* (very small) overhead is *fully absent* in the main GA cycles. In the cycles, it is known then that the list of needed leading addresses will not change anymore, *i.e.*, no additional lock is necessary any longer. Because of the loosely coupled overall design of such OO classes in OGOLEM and the clear structure on the overall (highest abstraction) execution flow, such a simple hook with encapsulated state and the address-space management system in terms of the object pool distribution could simply be added. At will (although there would be no need to), this could (in principle) simply be switched off without breaking the code and this would enable simple non-cached sPvREMD calls again.¹¹

Lastly, as future to-do, a similar scheme could also be implemented in the (generic)

¹⁰ At will one could, but should not, of course.

¹¹ That is, there is no input option or something similar available; but this is supposed to simply point to the fact that such important memory handling improvements by low-end explicit pointer manipulations can be added without the possibility at all to introduce subtle bugs because of the clear responsibilities of program parts, encapsulations, etc.

program workflow for **MPP** using the **RMI**-based parallelism.^[404,405] Here, between each process on possibly heterogeneous hardware, a thread-pool with shared memory can be triggered. Up to date, a related distribution of local addresses was not implemented, since the absolute run-times of REAXFF parameter optimizations seemed still to be manageable by **SMP** alone. Yet, as such an **OO** framework is set, this should result in an easy implementation.

Publication: Lennard–Jones Cluster Optimization

5.1 Scope of the Project

Lennard-Jones clusters^[167,174,180] constitute one of the standard benchmark systems for hard structural optimization problems in the realm of **cluster structure optimization**.^[164,165,397,448] Some sizes of these atomic clusters are known to be especially hard since the actual energetic global minimum lies far apart from most of the other minima on a rugged and deceptive landscape where the global minimum can be of a very different structural pattern (symmetry) than most of the other minima. Consequently, without any further algorithmic intrinsics to dodge such misdirection, many non-deterministic metaheuristics are prone to be trapped in these fallacious minima without finding the global one. Even more so, also recent developments in the literature with seemingly appealing new metaheuristics even have to confess to be unable to solve such well-known hard benchmark cases. In this context, we investigate some of these hard cases again by using our **EA**-based optimization in **OGOLEM** with focus on the additional order parameters, *i.e.*, niching,^[167] incorporating structural information of the clusters. In this way, diversity within the population is conserved (for a longer time) during the optimization to avoid being trapped. Two different types of protocols are used: On the one hand, a specific binning technique is introduced for recognizing sub-motifs in the clusters, which is intentionally a very problem-specific descriptor. On the other hand, we leverage global and abstract representations as used in recent **machine learning** research (*cf.* Section 2.6). In both cases, we achieved the needed guidance of the search for the diversion into the most promising region, the global minimum funnel. In fact, no over-specific fine-tuning for the problem is needed as long as *some* order parameter is used to hinder the premature convergence.

At the same time by capitalizing these benchmark problems, this publication served as a “touchstone” for the relational or dynamic niching and the **coulomb matrix** concepts that can equally be used in the **GOCAT** design (*cf.* Section 3.5.3).

5.2 Publication Data and Reprint

<i>Reference:</i>	M. DITTNER, B. HARTKE, CONQUERING THE HARD CASES OF LENNARD-JONES CLUSTERS WITH SIMPLE RECIPES, <i>Comput. Theor. Chem.</i> 2017 , <i>1107</i> , 7–13, DOI: 10.1016/j.comptc.2016.09.032. ^[449]
<i>Submitted:</i>	September 1, 2016.
<i>Accepted:</i>	September 26, 2016.
<i>Contribution:</i>	Implementation of niching based on variants of the coulomb matrix in OGOLEM, major parts of calculations (re-done for consistency for all niching types), analyses and discussions, major contribution to the text.
<i>Graphic:</i>	Illustrated in Fig. 5.1.
<i>ESI:</i>	Printed on pp. 294–298 in Appendix B.
<i>Copyright:</i>	Reproduced with permission from “Computational and Theoretical Chemistry”. Copyright 2016 Elsevier B.V.

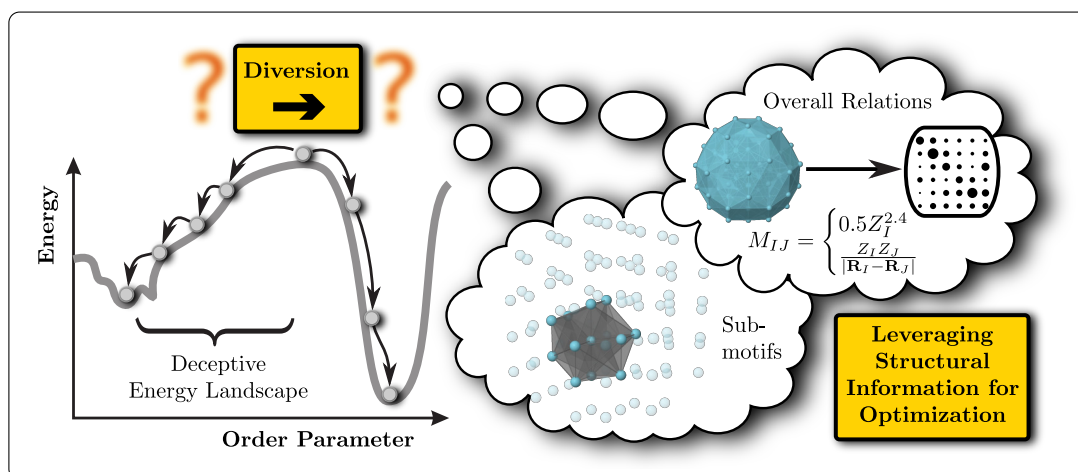


Fig. 5.1: Original “Table of Contents” graphic.



Contents lists available at ScienceDirect

Computational and Theoretical Chemistry

journal homepage: www.elsevier.com/locate/comptc

Conquering the hard cases of Lennard-Jones clusters with simple recipes



Mark Dittner, Bernd Hartke*

Institute for Physical Chemistry, University of Kiel, Olshausenstr. 40, 24098 Kiel, Germany

ARTICLE INFO

Article history:

Received 1 September 2016
 Received in revised form 24 September 2016
 Accepted 26 September 2016
 Available online 28 September 2016

Keywords:

Non-deterministic global optimization
 Evolutionary algorithms
 Cluster structures
 Order parameters
 Deceptive energy landscapes

ABSTRACT

Lennard-Jones clusters are the best-known benchmark for global cluster structure optimization. For a few cluster sizes, the landscape is deceptive, featuring several funnels, with the global minimum not being in the widest one. More than a decade ago, several non-deterministic global search algorithms were presented that could solve these cases, mostly using additional tools to ensure structural diversity. Recently, however, many publications have advertised new search algorithms, claiming efficiency but being unable to solve these harder benchmark cases. Here, we demonstrate that evolutionary algorithms can solve these hard cases efficiently, if enhanced with one of several very different diversity measures (niching) which were set up in an ad-hoc way, without extensive deliberation, testing or tuning. Hence, these hard benchmark cases should definitely be considered solvable. Additionally, these niching concepts offer insights into the different Lennard-Jones structural types, and into the way niching works in evolutionary algorithms.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Clusters of n atoms bound exclusively by pairwise Lennard-Jones (LJ) model potentials of the form $\tilde{E} = 4\sum_{i<j}(\tilde{r}_{ij}^{-12} - \tilde{r}_{ij}^{-6})$ (with the energy E and the pair distance r in reduced units of the pair well depth and distance) have been a standard benchmark for global cluster structure optimization algorithms for a long time [1–17]. As frequently noted in these and other studies, the global minima for most of the cluster sizes n are Mackay icosahedra and easy to find, despite the exponential increase of search space size with n . However, for a few isolated cluster sizes, the structure of the true global minima is different: decahedral for $n = 75, 76, 77, 102, 103, 104$, face-centered cubic (fcc) for $n = 38$, and tetrahedral for $n = 98$. Astonishingly, the latter case was discovered only in 1999 by Leary and Doye [18], i.e., it was missed by several of the first studies cited above, which documents that it is hard to find.

These isolated occurrences of different global minimum structures are linked to partially filled structural shells and the different ways structural strain (deviations from ideal pair distances) can be accommodated in different structural types, as clearly illustrated by Doye et al. [19]. These authors also demonstrated that locating the true global minima in these cases is hard because most of search space is still dominated by the standard icosahedral pattern and its associated funnel-like landscape, while the different struc-

tural patterns (also containing the global minimum) reside only in a small region of search space, isolated from the remainder by high energy barriers.

This also explains why the case $n = 38$ was considered very hard in the early days of LJ cluster studies, despite its small size, and why it is apparently acceptable to admit problems with the larger hard cases in publications up to the present day, despite explorations towards sizes up to $n = 1000$ and beyond quite some time ago [20]. For example, Lv et al. [10] reported good results for $n = 75$ but failed to find the T_d minimum for $n = 98$ in 7 out of 10 cases. Laykhov et al. [11] called $n = 75$ “exceptionally complex”. Rogan et al. [12], Zhang et al. [16] and Avendaño-Franco et al. [17] even failed to find the decahedral global minimum for $n = 75$, the latter two in publications of the present year 2016. This is astonishing, given that 10–15 years earlier, several publications, e.g., Refs. [3,21,22], had already presented recipes that successfully reduced the search effort for several or all of these hard cases. Therefore, the present contribution serves to reconfirm those earlier works: Present-day publications aspiring to conform to the state of the art should be able to deal with these hard cases, because this does not require specialized, fine-tuned recipes but merely a somewhat more judicious and robust design of the search algorithm.

In fact, what is needed has been known since the early days of non-deterministic global search and has been re-analyzed many times, also in recent years [23]: The practical strength of these algorithms lies in their deliberate refusal to cover all search space; instead the search is narrowed down on “promising regions”. This

* Corresponding author.

E-mail address: hartke@pctc.uni-kiel.de (B. Hartke).<http://dx.doi.org/10.1016/j.comptc.2016.09.032>

2210-271X/© 2016 Elsevier B.V. All rights reserved.

can lead to very large performance enhancements, compared to deterministic search, which always has to cover all search space, at least indirectly. In many cases in practice, this makes the difference between being able to solve a global optimization problem and having to give up. The price to pay for this advantage is that so-called “deceptive” search landscapes can trap non-deterministic search in regions that do not contain the global minimum. Hence, mechanisms are needed to avoid such a trapping.

Obviously, for LJ clusters, the trick is to avoid that all search power is spent within the broad icosahedral basin, which acts as a strong attractor for non-deterministic search. This has been done already, for example with niching in Evolutionary Algorithms (EAs) [3]. As a side note, we are very much in favor of EA nomenclature to become “less inspired” [24,25]; in this sense “niching” should be called “introduction of an order parameter” instead. Nevertheless, to make contact to previous EA literature, we continue to use the biologically inspired term “niching” here. In Ref. [3], structures similar to the icosahedral and decahedral type were differentiated by rotating each cluster into an orientation in which a two-dimensional plane projection of its atom positions was least dense, and then calculating this density as the fraction of occupied squares in a discretization of this plane. Icosahedral structures have a significantly higher projected density than decahedral ones. The actual niching then allows only a small number of individuals (much smaller than the whole population) to have similar projected densities. This projection niching in Ref. [3] was very much ad-hoc, tainted with a priori knowledge, and computationally expensive, since the desired differentiation can only be made very close to the ideal cluster orientation, requiring a long sequence of small incremental test rotations, at each of which the 2D projection has to be evaluated.

Within their adaptive immune optimization algorithm (AIOA), Cheng et al. [21] have based their niching-like diversity concept on differences in nearest-neighbor connectivity table entries, between two structures. This depends on the proper choice of a cutoff criterion, to discern small differences in nearest-neighbor distances. Otherwise, with a looser cutoff criterion, all inner atoms always have 12 nearest neighbors, as shown below and as to be expected for closest packings between particles with non-directional interactions. However, with proper choice of this cutoff, these authors achieved impressive efficiency for the LJ hard cases, including $n = 98$.

Rossi and Ferrando [22] implemented a similar niching-like concept in Monte Carlo with Minimization (MCM) [26], also known as basin-hopping (BH) [27,28]. In their implementation, several simultaneous MCM walkers repel each other in an order-parameter space. With suitable choice of these order parameters, exploration can be diverted into different funnels. For LJ clusters, they found significant search efficiency enhancement for $n = 38$ and 75. To differentiate between icosahedral, decahedral and fcc structures, they chose the common neighbor analysis (CNA) [29–32].

CNA is one of several ways [33] to categorize nearest neighbor arrangements of atoms. It is used frequently to detect structural faults, domain boundaries and phase transitions in bulk MD simulations [34,35], but also for structural characterization of clusters [36,37]. In CNA, to each atom pair, an integer triple (m, n, k) is assigned, with m nearest neighbors common to both atoms in the pair, between which there are n bonds, and k bonds of these form the longest connected chain. As pointed out by Ferrando et al. [22,38,39], it is sufficient to monitor the CNA signatures (5,5,5), (4,2,2) and (4,2,1) to distinguish icosahedral-, decahedral- and fcc-structured clusters.

While Rossi and Ferrando have shown [22] that CNA-based differentiation does help for the LJ hard cases $n = 38$ and 75, it is

unclear if it also works for $n = 98$ with the different T_d structure. Further possible downsides of CNA are that it is pair-based instead of atom-centered, and that intuitive correspondences between the (m, n, k) designation and actual local neighborhood structures are unclear (except for (5,5,5) which is normally linked to local 5-fold symmetry axes).

To emphasize with the present contribution that special characteristics of niching or diversity concepts are not important and that LJ hard cases can be solved by essentially any reasonable concept of this kind, we present two nichings that do have some aspects of similarity with the earlier ones but also several differences, and, most importantly, strongly differ from each other. Nevertheless, they achieve similar degrees of efficiency, when compared with each other and with earlier results, as mentioned above.

The first niching concept is based on a different local neighborhood categorization, which is atom-centered, can also differentiate T_d from icosahedral, decahedral and fcc, and is intuitively understandable. Hence, it also contributes insights into how these four basic LJ structural types differ, at the level of local nearest-neighbor arrangements. When this categorization is used to define niches in an evolutionary algorithm, this enables the EA to solve all these hard LJ cases with one and the same setting.

The other niching concept is based on the so-called Coulomb matrix (CM in the following), \mathbf{M} , which is used, e.g., also in Machine Learning studies as common measure of similarity throughout the chemical compound space (see Refs. [40,41] and references therein for a small overview as well as restrictions of the measure used here and Ref. [42] as a general investigation on similarity measures). Any cluster thus is represented by

$$M_{ij} = \begin{cases} 0.5Z_i^{2.4} & \text{for } I = J, \\ \frac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} & \text{for } I \neq J, \end{cases} \quad (1)$$

with the atomic nuclear charges Z and distances \mathbf{R} between the atoms I and J . The CM represents the Coulomb repulsion on the non-diagonal elements and a polynomial fit of the nuclear charges to the total energies of free atoms on the diagonal ones [43,44]. Note that in the case of non-mixed (atomic) LJ-clusters of this paper, all nuclear charges are the same, such that this reduces effectively to a matrix storing all N^2 (redundant) distances between the atoms. To construct a (dis-)similarity measure between two clusters, we use the Euclidean norm of the diagonalized CMs:

$d(\mathbf{M}, \mathbf{M}') = d(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}') = \sqrt{\sum_i |\epsilon_i - \epsilon'_i|^2}$ with $\boldsymbol{\epsilon}$ as ordered eigenvalues of \mathbf{M} . This now represents a translation-, rotation- and (atomic) permutation-invariant measure. Note, though, that this descriptor is not unique but coarsened, as in the eigenvalue vectors of the clusters effectively only N items of information are included (the additional information carried by the eigenvectors is completely discarded). However, as we want to form coarsened similarity niches over the energy landscape, this is in our case no disadvantage. Additionally, with a very small threshold on d , this can even be used in most cases as identity check for LJ clusters of the sizes studied here, despite the non-uniqueness.

2. Generic niching implementations

Global cluster structure optimization has been done here with the universal, object-oriented `OGOLEM` package [45,46], which has already been applied to a wide variety [9,47–53] of global optimization tasks. To avoid serious serial bottlenecks, `OGOLEM` implements the generation-free pool concept [54]. A generic niching implementation has been integrated with the pool concept allowing for arbitrary niching criteria to be employed.

2.1. Deterministic mapping

As first generic niching framework, a “static” binning (i.e., a *classification*) algorithm is used by deterministically assigning a niche ID to a given individual. The responsibility to provide an adequate resolution of these niches, i.e., how broadly a single niche is defined, lies with the niching criterion implementation. The niche ID itself is encoded as a string of characters of arbitrary length. The generic niching backend keeps track of how many individuals share the same niching ID. If the population of any niching ID exceeds a user-defined threshold, the backend rejects solution candidates added to the pool if their fitness is less favorable than the worst individual in their niche even if it is better than the worst individual in the overall pool. If the candidate's fitness is better than the one of the worst individual in the same niche, the new candidate is added to the pool and the worst individual in the same niche is removed from the pool.

Assuming a niching criterion implementation that successfully differentiates between relevant candidate properties, this procedure significantly reduces the possibility that premature convergence of the genetic pool with only a reduced, wrong set of genomic properties occurs. At any give time, at least

$$N_{\text{niches}} = \frac{\text{pool size}}{\text{max. niche population}} \quad (2)$$

will be present in the genetic pool, each representing a unique property as defined by the niching criterion. We can therefore summarize the demands for the given niching criterion as follows:

- provide deterministic mapping of individual's genome to a niche ID,
- identify relevant genome properties and represent them in the ID,
- provide sufficiently unique and sufficiently broad definition of niche IDs.

The first niching type shown in this article, Section 3, which is based on local neighborhoods, implements this ID mapping scheme.

2.2. Coarsened similarity measure

A second generic niching framework can use any arbitrary similarity descriptor between two individuals. I.e., in the deterministic mapping above, a unary operator assigns (classifies) one cluster statically to one niche. In contrast, in this scheme we instead use a “dynamic” niching: Each individual that is born is compared by a binary operator to the rest of the population. When the computed similarity measure – in the following this will be the CM-measure $d(\epsilon, \epsilon')$ (cf. Section 4) – between two individuals is less than a user-supplied threshold value, the individuals are assigned to the same niche. Thus, this can also be interpreted more like an agglomerative clustering algorithm, where similar individuals dynamically define their own similarity “clouds”, i.e. the niches, during optimization. Still, the general framework of adding and/or removing niches during the global optimization stays the same as mentioned above. Again, the important requirement is to chose a meaningful similarity measure which is able to encode the characteristic details of the genotypes of individuals and decide on being *similar* or *different* during *comparison* without pre-assigned (classified) niches. Note that we follow not a complete current population clustering scheme in every iteration of the global optimization. Because of our always fitness-sorted population and elitism, we include a “bias” of preferentially comparing new incoming individuals starting at the better ones in the population. This way, the

currently best individuals (maybe forming one basin of attraction) are prioritized to form a cloud (which may not be over-populated because of the niching), while the worst individuals are compared less often, only if no similar individual was found among the better ones.

3. Niching based on local neighborhoods

The generic niching implementation discussed above (cf. Section 2.1) entails that classifying new clusters into niches is done after every crossover/mutation step. Therefore, using the projection niching cited above [3] is too costly; in that previous publication it was affordable because it was done only on the best individuals already selected for survival into the next generation.

However, during our testing of possible alternative niche definitions, it turned out that all low-energy LJ clusters contain only four nearest-neighbor configurations around non-surface atoms, as depicted in Fig. 1. The slightly non-standard names assigned to these four configurations emphasize that there are actually only two different ones (icosahedral and fcc), with two subtypes each (staggered or eclipsed arrangement of the top and bottom parts relative to each other). All of these feature 12 nearest neighbors, as in typical densest sphere packings, and are familiar building blocks of crystal structures, so they could have been guessed in advance.

These four nearest-neighbor configurations can be discerned easily by different means; in this work, the numbers of (approximate) right angles and of longest distances (within a margin of 5%) among the 12 atoms surrounding the central one were counted, yielding a characteristic result for each type.

Note that these four types do not agree with the CNA triples mentioned above, and cannot be cleanly differentiated by them. The CNA-triple (5,5,5) is the only one present in ico-staggered, but it occurs as well in ico-eclipsed (for 16.7% of the pairs). The majority of pairs in ico-eclipsed is (4,2,2), which also characterizes half of the pairs in cp-ABA. The other half of cp-ABA is (4,2,1), which occurs exclusively in cp-ABC. In this sense, these four nearest-neighbor configurations correspond better and more intuitively to actual atom-centered surroundings.

Furthermore, these four nearest-neighbor configurations (NC) occur with characteristically differing percentages in low-energy, real-life LJ structures, cf. Table 1. The percentages (and their ranges) given in this table were established by comparing the output percentages from our classification algorithm to overall cluster structure types that were obtained by visual inspection, for a test set of three dozens of clusters, containing putative global minima, and several low-energy minima. It was verified that these percentages and percentage ranges are sufficiently accurate by again comparing automatic and visual inspection, for several new global optimization runs for differing cluster sizes.

This allows to set up a selective NC-niching in two different ways: In *mode 1*, the nearest-neighbor percentage distributions of Table 1 are known a priori, and are used (with suitable deviation margins) to define niches for the overall structural types ico, deca, T_d and fcc; with ico as default niche for all non-classifiable structures (which coincides with the observation that most structures in LJ search space are ico anyway). Clearly, this constitutes a considerable amount of a priori knowledge. In contrast, in *mode 2*, the only a priori knowledge used are the four nearest-neighbor configurations shown in Fig. 1, and niches are defined by four numbers, each of them a binned percentage of these four nearest-neighbor configurations occurring in the cluster in question, with a bin width of 20 percentage points. Obviously, this allows for many more niches than in *mode 1*. To accommodate that, pool sizes for *mode 2* should be significantly larger.

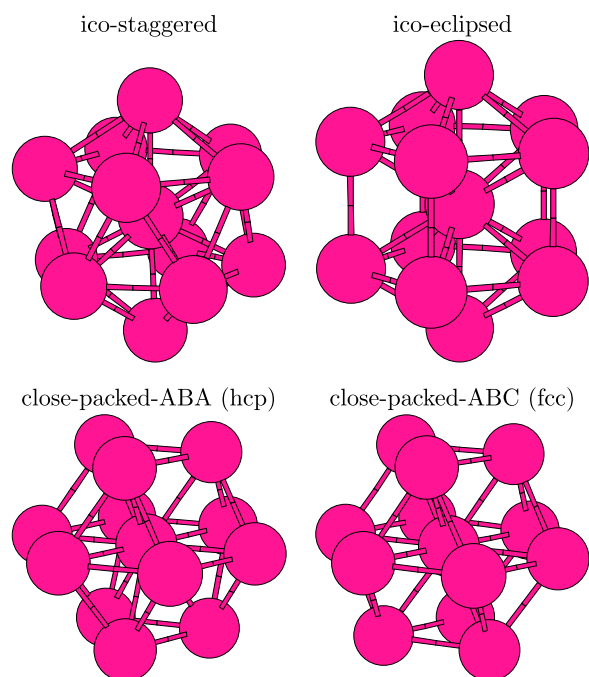


Fig. 1. The four nearest-neighbor configurations found inside low-energy LJ clusters.

Both niching modes markedly enhance the ability of EA search to find LJ global minima, in particular for the infamous difficult cases. Of course, *mode 1* with its greater amount of a priori information provides a larger boost, but *mode 2* is not only possible, too, but also constitutes a big gain in performance. To put this gain in perspective, let us emphasize that repeating the runs below without niching (but with all other settings unchanged) is highly unsuccessful: With very few lucky exceptions, none of the known global minima are found within 20 million global steps, despite using non-standard moves like our graphical directed mutation [55].

The second column in Table 2 and the corresponding curves in Figs. 2 and 3 illustrate EA efficiency with *mode 1* niching. (For comparison with literature results, note that global iteration step numbers roughly correspond to numbers of local optimizations; since our graph-based directed mutation [55] also includes some limited local optimization cycles, this correspondence is not perfect. Also note that this Table and these Figures collect all results, including those in the remainder of this Article, for easy comparison. Separate Tables for each case, including numbers for standard deviations and earliest/latest encounters, can be found in the Supporting Information.) With a poolsize of 100 individuals, at most 20 individuals were allowed for the five niches ico, deca, T_d , fcc and mixed, so that an even distribution over all niches was possible. However, of course, no direct or indirect forcing towards

these structural types was used. Instead, starting from random structures, standard EA crossover and mutation operators were employed, together with the graph-based directed mutation [55] introduced earlier. As comparison to a non-hard case of a similar size, the last row of Table 2 also shows results for $n = 100$, generated with the exact same settings. There, the accepted global minimum is of the ico type, but due to the half-filled outer shell, the search landscape can be expected to be more challenging than for the very strongly ico-dominated, closed-shell cases $n = 55$ and $n = 147$.

Clearly, *mode 1* niching makes it possible to find the accepted global minima for all of these hard cases in well under 3 million global optimization steps, or well under 500 thousand steps on average. This is to be contrasted with attempting the same runs with the same settings but with niching switched off: Then, only in rare lucky cases, earliest encounters of global minima for these hard cases occur before 20 million steps, indicating that average first encounters occur well beyond 20 million steps. With *mode 1* niching, average step numbers until the first encounter still are greater than for simpler cases like $n = 100$, but only by a factor between 1 and 10 – with the exception of $n = 98$, which obviously is harder still, due to two elusive structural types (deca and T_d) being competitive in energy with the dominant ico type. However, earliest encounters are very similar for all cases, indicating that niching does what it is supposed to do, namely protecting not-yet-perfect deca and T_d structures from being outperformed by ico structures. In contrast, latest encounters are much later for the hard cases than for $n = 100$, reflecting the same differences as for the average first encounters.

The third column in Table 2 compares this to EA efficiency with *mode 2* niching (also see Figs. 2 and 3). As indicated, this mode makes many more niches possible, hence a larger pool of 2000 individuals was used. Again, a maximum of 20 individuals per niche was allowed.

Comparing niching *mode 1* and 2 for average first encounters, no clear trend emerges; sometimes *mode 1* is better (for $n = 98$ and 104), sometimes *mode 2* is better (for $n = 76, 77$ and 103), and the numbers are similar for the remaining cases. What differs more consistently are the earliest encounters: With *mode 2*, they always take about one order of magnitude longer. This can be rationalized by the presence of more niches, which also protect intermediate structural types that differ from the optimal ones. This also explains why *mode 2* niching makes the search for not-so-difficult cases like $n = 100$ less efficient. Nevertheless, it is surprising that the strongly reduced amount of a priori knowledge is not more damaging: For the hard cases, *mode 2* niching is very beneficial, and overall just as good as *mode 1*.

4. Niching based on the Coulomb matrix

According to the general details about the Coulomb matrix of Section 1, we now follow the second niching implementation based on an arbitrary similarity measure (cf. Section 2.2), i.e., $d(\epsilon, \epsilon')$ as measure between two individuals.

The fourth column in Table 2 presents EA efficiency with CM-niching in *CM setting 1* (also see Figs. 2 and 3). Similar to NC-

Table 1
Percentages of the four nearest-neighbor configurations (Fig. 1) occurring inside low-energy LJ clusters.

Cluster	ico-staggered	ico-eclipsed	cp-ABA	cp-ABC
ico	1–2 (count)	50–60%	40–50%	0
deca	0	10–17%	48–54%	18–33%
T_d	0	43%	57%	0
fcc(98)	0	0	43%	38%
fcc(38)	0	0	0	100%

Table 2

Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and several niching concepts, averaged over 20 runs each; in thousands of steps, and rounded to 2 significant digits.

Cluster size	NC niching		CM niching		Global best
	Mode 1	Mode 2	Setting 1	Setting 2	
75	63	41	180	120	39 ^a
76	220	80	290	320	80 ^b
77	320	94	900	310	94 ^b
98	450	630	60	220	36 ^c
102	32	34	51	53	32 ^d
103	48	24	60	61	24 ^e
104	48	240	140	130	41 ^e
100	24	130	40	57	

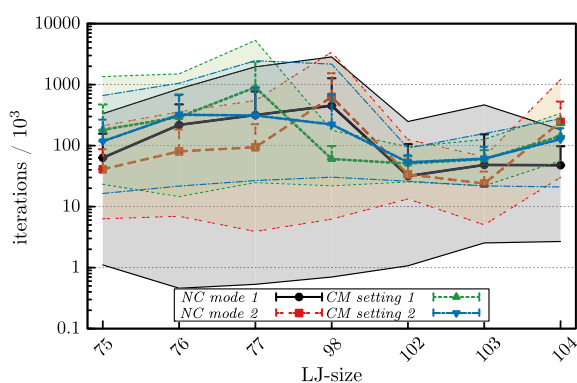
^a NC mode 4: equal to mode 2 (see text), but using a smaller bin width of 10 percentage points.

^b NC mode 2: explained in main text.

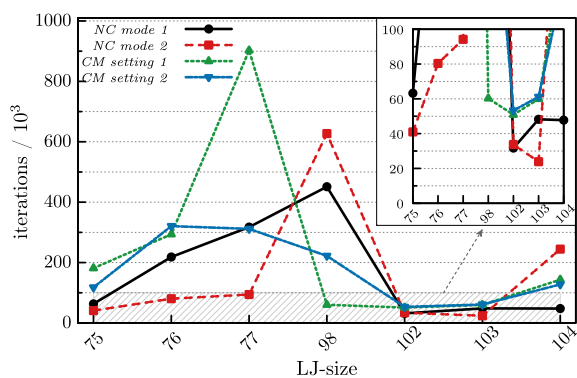
^c CM setting 3: using smaller $d(\mathbf{M}, \mathbf{M}') = 5$.

^d NC mode 1: explained in main text.

^e NC mode 3: equal to mode 1 (see text), but using a smaller bin width of 10 percentage points.



(a) Max. & min. (filled curve), standard deviation (just upper half error bar) and average iterations.



(b) Average iterations (zoom).

Fig. 2. Average global optimization steps for the first encounter of the true global minimum, for several cases (average iterations compiled in Table 2; all other details given in the Supporting Information).

niching in mode 2, we used a bigger pool of 2000 individuals, and at most 10 individuals in one niche. In preliminary studies, we examined various threshold values, $d_{\text{thresh}}(\epsilon, \epsilon')$, to be able to differentiate most meaningful local and global minima in test runs. A threshold of 10 was used¹ in the results shown here, if not stated otherwise (however, below we will discuss how the character of

¹ For CM construction, Eq. (1), we used atomic units throughout.

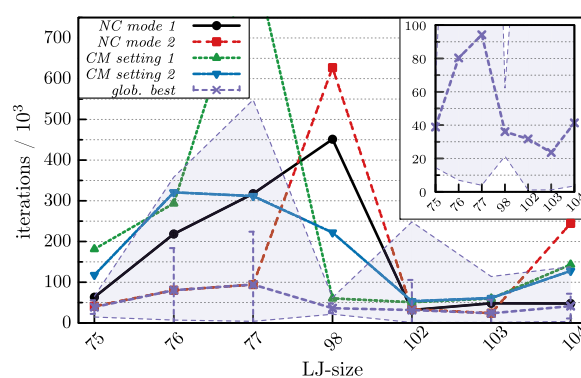


Fig. 3. Average global optimization steps for the first encounter of the true global minimum, for the cases of Fig. 2 together with the “globally best” collection (cf. Table 2 and the Supporting Information for more details about the specific niching types used): Mixed results of mode 1 & 2 (also including a width of 10, not 20), and CM (also including different threshold parameters). For the latter, the max. (latest), min. (earliest) encounter is plotted (filled curve), standard deviation (error bar) and average iterations.

CM-niching can be changed by choosing other values). All other GA-settings like GA-operators were the same as in the optimizations above (mode 1 & 2 in Section 3).

The most striking observation at this stage is that the $n = 98$ case now is exceptionally easy with up to one order of magnitude less iterations than with NC-niching. However, this benefit is at the expense of performance penalties at $n = 75, 76, 77$ clusters, which will be discussed below.

Compared to NC-niching, CM-niching definitely is a more abstract version of a cluster similarity measure, without clear heuristics. Therefore, we tried several different CM and EA settings to probe its possible overall performance. As one example (setting 2), in the fifth column of Table 2 the results of a CM-niching is shown that is a compromise between the performance at $n = 75, 76, 77$ vs. $n = 98$, with similar performance at the other cluster sizes shown. Compared to the NC-niching shown above, in this case we have used less fitness pressure by the selection operator: One random individual per niche is selected while the other one belongs to the better individuals in that niche. Also, the additional energy diversity check per niche was increased from 0.01 (used in all other cases shown) to 0.05 as a percentage of energy deviation between structures to be treated as being different.

With this setting, it is possible to reach all the global minima in about (or less than) 300 k iterations on average.

Note that the threshold of CM-similarity strongly influences overall performance: For example, by shrinking the threshold to $d_{\text{thresh}}(\epsilon, \epsilon') = 5$, the character of CM-niching could be changed to conserve even more diversity, such that we can observe an overall performance of 36 k iterations for reaching the global T_d minimum in the $n = 98$ case. In another case, we could improve the best performance of our CM-niching for $n = 75, 76, 77$ by using a bigger threshold of $d_{\text{thresh}}(\epsilon, \epsilon') = 15$. This results in broader niches with more differing structures in it (approaching some of the character of NC-niching). With this setting, the global minimum of $n = 77$ is reached in 140 k iterations ($n = 75, 76$ in about 100 k iterations), but we suffer a penalty for $n = 98$, which then needs more than 1 million iterations.

In fact, of course it is possible to fine-tune these niching recipes, for specific cluster sizes. Results from a few steps in this direction are given in the last column of Table 2 and in Fig. 3, which also contains results mentioned in the previous paragraph. However, for new real-life applications, compromise settings that robustly provide fairly good performance across many cases are more important. Those are the settings we have presented and discussed above.

Obviously, CM-niching and NC-niching have rather different impacts on the EA search: With NC-niching, $n = 98$ clearly is the most difficult case, while $n = 75, 76, 77$ are less difficult, and $n = 102, 103, 104$ only are a modest challenge. In contrast, CM-niching is most successful for $n = 98$, with the other cases frequently being harder (depending on the other EA settings).

In addition, one may ask why CM-niching works as well as it does, namely essentially on par with NC-niching, if used with one compromise setting across all the hard cases. As explained in the previous section, NC-niching is based on structural insights specific for LJ clusters – even in *mode 2*, where these insights are not used to pre-set known niches but only to define the features used for structural differentiation. In contrast, CM-niching knows nothing about LJ clusters. In fact, it does not even know anything about clusters: It takes all atoms into account, whereas in NC-niching the outer shell is stripped off as a first step, since it can be expected to be distorted away from optimality due to surface reconstructions.

It turns out that all these observations presumably are linked to each other, according to our working hypotheses on how NC- and CM-niching work. In CM-niching, indeed every atom counts, therefore structures that differ only in positions of a few atoms are likely to end up in different niches. In contrast, structures in the same NC-niche may and do differ in the positions of comparatively many atoms but still share the same type of buildup in their cores. Hence, CM-niching supports a higher degree of exploration or leads to a lower selection pressure. Apparently, this is beneficial for $n = 98$, where not just two but three major structural forms (ico, deca, T_d) are in close competition for the global minimum. With niches even more spread out in search space, chances to discover the T_d global minimum are much increased.

This advantage of CM-niching for $n = 98$ may in turn be its disadvantage for $n = 75, 76, 77$: There, the decahedral global minima have a markedly oblate outer shape, while the best icosahedral structures are closer to being spherical or prolate. This outer shape difference is much less pronounced for $n = 102, 103, 104$. Hence, and since our moveclass does not contain moves specifically designed to change outer shape while preserving inner structure, $n = 102, 103, 104$ is simpler than $n = 75, 76, 77$ for both kinds of niching. Nevertheless, NC-niching can deal better with this problem due to its wider niches: Our crossover and mutation operators have a reasonable chance to transform two already decahedral structures into two better decahedral structures, even within one and the same niche. With CM-niching, however, the path to the

oblate decahedral global minimum is more similar to a discovery from scratch.

5. Conclusions

We have shown here that proper diversity ingredients are essential to enable non-deterministic global search to cope with deceptive search landscapes, exemplified here by the famous LJ hard cases $n = 75, 76, 77, 98, 102, 103, 104$. This necessity was well known in the early days of non-deterministic search (applied to cluster structures [3] but also in general). However, awareness of this issue appears to have decreased in recent years, as witnessed by recent papers that advertise their variants of search algorithms as powerful and efficient but openly admit strong difficulties or even outright failures for these LJ hard cases.

Hence, we have demonstrated that fine details of diversity tools, here in the form of niching in an EA, are not very important: With two very different niching concepts we could achieve essentially similar overall performance (with differences in performance for different cluster sizes that are to be expected, given that search space structures do depend on cluster size, and which can be explained). Also, the level of efficiency is essentially the same as that obtained previously, with diversity tools that again differ in their implementation details: For example, Chen et al. [21] reported mean first encounters of the accepted global minima for $n = 75, 98, 102$ in the range of 30,000–40,000 local minimization steps, which matches the best results we obtain.

In addition, our NC-niching offers additional understanding of the inner structure of these clusters. In contrast, our CM-niching is more abstract but uses a tool that has found widespread use recently in the machine-learning community.

Since soft sphere packing effects are present even in molecular clusters with directional bonding [56] and since we have shown that niching/diversity details are less important than expected, we are confident that essential parts of the recipes presented here can be generalized from the LJ cluster benchmark to real-life systems. And we hope that the present work convinces other users of the LJ cluster benchmark that failures in dealing with the known hard cases are not acceptable anymore and can be remedied easily.

Acknowledgements

It is a pleasure for us to thank Johannes Dieterich for daring us to try NC niching *mode 2*, but far more importantly for starting to build the marvelous all-purpose global optimization suite `OGOLEM` during his PhD in the Hartke group, and for continuing its development ever since, with amazing intensity and enthusiasm. This has made not only the present work possible and easy, but also many other projects in very diverse areas.

We thank Hannes Jónsson for sending us his original papers on CNA. BXH thanks the German Research Foundation DFG for financial support of this work, through Grant Ha2498/16-1. MXD gratefully acknowledges financial support by a fellowship of the German Fonds of the Chemical Industry (FCI).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.comptc.2016.09.032>. These data include more detailed information on the tests mentioned in the main article (pdf file), all tabulated numbers (csv files), and atomic cartesian coordinates of the most important structures (pdb files).

References

- [1] D.M. Deaven, N. Tit, J.R. Morris, K.M. Ho, Chem. Phys. Lett. 256 (1996) 195, [http://dx.doi.org/10.1016/0009-2614\(96\)00406-X](http://dx.doi.org/10.1016/0009-2614(96)00406-X).
- [2] D.J. Wales, J.P.K. Doye, J. Phys. Chem. A 101 (1997) 5111, <http://dx.doi.org/10.1021/jp970984n>.
- [3] B. Hartke, J. Comput. Chem. 20 (1999) 1752, [http://dx.doi.org/10.1002/\(SICI\)1096-987X\(199912\)20:16<1752::AID-JCC7>3.0.CO;2-0](http://dx.doi.org/10.1002/(SICI)1096-987X(199912)20:16<1752::AID-JCC7>3.0.CO;2-0).
- [4] C. Barrón, S. Gómez, D. Romero, A. Saavedra, Appl. Math. Lett. 12 (1999) 85, [http://dx.doi.org/10.1016/S0893-9659\(99\)00106-8](http://dx.doi.org/10.1016/S0893-9659(99)00106-8).
- [5] X. Shao, H. Jiang, W. Cai, J. Chem. Inform. Comput. Sci. 44 (2004) 193, <http://dx.doi.org/10.1021/ci0340862>.
- [6] X. Shao, L. Cheng, W. Cai, J. Comput. Chem. 25 (2004) 1693, <http://dx.doi.org/10.1002/jcc.20096>.
- [7] W. Pullan, J. Comput. Chem. 26 (2005) 899, <http://dx.doi.org/10.1002/jcc.20226>.
- [8] H. Takeuchi, J. Chem. Inform. Model. 46 (2006) 2066, <http://dx.doi.org/10.1021/ci600206k>.
- [9] J.M. Dieterich, B. Hartke, J. Comput. Chem. 32 (2011) 1377, <http://dx.doi.org/10.1002/jcc.21721>.
- [10] J. Lv, Y. Wang, L. Zhu, Y. Ma, J. Chem. Phys. 137 (2012) 084104, <http://dx.doi.org/10.1063/1.4746757>.
- [11] A.O. Lyakhov, A.R. Oganov, H.T. Stokes, Q. Zhu, Comput. Phys. Commun. 184 (2013) 1172, <http://dx.doi.org/10.1016/j.cpc.2012.12.009>.
- [12] J. Rogan, A. Varas, J.A. Valdivia, M. Kiwi, J. Comput. Chem. 34 (2013) 2548, <http://dx.doi.org/10.1002/jcc.23419>.
- [13] M.T. Oakley, R.L. Johnston, D.J. Wales, Phys. Chem. Chem. Phys. 15 (2013) 3965, <http://dx.doi.org/10.1039/c3cp44332a>.
- [14] H. Takeuchi, Comput. Theor. Chem. 1050 (2014) 68, <http://dx.doi.org/10.1016/j.comptc.2014.10.017>.
- [15] H. Takeuchi, Chem. Phys. 457 (2015) 106, <http://dx.doi.org/10.1016/j.chemphys.2015.05.026>.
- [16] J. Zhang, M. Dolg, Phys. Chem. Chem. Phys. 18 (2016) 3003, <http://dx.doi.org/10.1039/C5CP04060D>.
- [17] G. Avendaño-Franco, A.H. Romero, J. Chem. Theory Comput. 12 (2016) 3416, <http://dx.doi.org/10.1021/acs.jctc.5b01157>.
- [18] R.H. Leary, J.P.K. Doye, Phys. Rev. E 60 (1999) R6320, <http://dx.doi.org/10.1103/PhysRevE.60.R6320>.
- [19] J.P.K. Doye, M.A. Miller, D.J. Wales, J. Chem. Phys. 111 (1999) 8417, <http://dx.doi.org/10.1063/1.480217>.
- [20] Y. Xiang, L. Cheng, W. Cai, X. Shao, J. Phys. Chem. A 108 (2004) 9516, <http://dx.doi.org/10.1021/jp047807o>.
- [21] L. Cheng, W. Cai, X. Shao, Chem. Phys. Lett. 389 (2004) 309, <http://dx.doi.org/10.1016/j.cplett.2004.03.125>.
- [22] G. Rossi, R. Ferrando, Chem. Phys. Lett. 423 (2006) 17, <http://dx.doi.org/10.1016/j.cplett.2006.03.003>.
- [23] T. Weise, R. Chiong, K. Tang, J. Comput. Sci. Technol. 27 (2012) 907, <http://dx.doi.org/10.1007/s11390-012-1274-4>.
- [24] T. Weise, Why Research in Computational Intelligence Should Be Less Inspired. <<http://www.it-weise.de/thoughts/text/eclnspiration.html>> (accessed: 2016-05-09).
- [25] K. Sørensen, Int. Trans. Oper. Res. 22 (1) (2015) 3–18, <http://dx.doi.org/10.1111/itor.12001>.
- [26] Z. Li, H.A. Scheraga, Proc. Natl. Acad. Sci. USA 84 (1987) 6611, <http://dx.doi.org/10.1073/pnas.84.19.6611>.
- [27] J.P.K. Doye, D.J. Wales, J. Phys. Chem. A 101 (1997) 5111, <http://dx.doi.org/10.1021/jp970984n>.
- [28] D.J. Wales, H.A. Scheraga, Science 285 (1999) 1368, <http://dx.doi.org/10.1126/science.285.5432.1368>.
- [29] H. Jónsson, H.C. Andersen, Phys. Rev. Lett. 60 (1988) 2295, <http://dx.doi.org/10.1103/PhysRevLett.60.2295>.
- [30] A.S. Clarke, H. Jónsson, Phys. Rev. E 47 (1993) 3975, <http://dx.doi.org/10.1103/PhysRevE.47.3975>.
- [31] D. Faken, H. Jónsson, Comput. Mater. Sci. 2 (1994) 279, [http://dx.doi.org/10.1016/0927-0256\(94\)90109-0](http://dx.doi.org/10.1016/0927-0256(94)90109-0).
- [32] E. Maras, O. Trushin, A. Stukowski, T. Ala-Nissila, H. Jónsson, Comput. Phys. Commun. 205 (2016) 13, <http://dx.doi.org/10.1016/j.cpc.2016.04.001>.
- [33] A. Stukowski, Model. Simul. Mater. Sci. Eng. 20 (2012) 045021, <http://dx.doi.org/10.1088/0965-0393/20/4/045021>.
- [34] P. Geiger, C. Dellago, J. Chem. Phys. 139 (2013) 164105, <http://dx.doi.org/10.1063/1.4825111>.
- [35] S.C. Kapfer, W. Mickel, K. Mecke, G.E. Schröder-Turk, Phys. Rev. E 85 (2012) 030301, <http://dx.doi.org/10.1103/PhysRevE.85.030301> (R).
- [36] H.U. Rehman, M. Springborg, Y. Dong, J. Phys. Chem. A 115 (2011) 2005, <http://dx.doi.org/10.1021/jp109198r>.
- [37] G. Rollmann, M.E. Gruner, A. Hucht, R. Meyer, P. Entel, M.L. Tiago, Phys. Rev. Lett. 99 (2007) 083402, <http://dx.doi.org/10.1103/PhysRevLett.99.083402>.
- [38] F. Baletto, C. Mottet, R. Ferrando, Phys. Rev. B 63 (2001) 155408, <http://dx.doi.org/10.1103/PhysRevB.63.155408>.
- [39] F. Baletto, R. Ferrando, Rev. Mod. Phys. 77 (2005) 371, <http://dx.doi.org/10.1103/RevModPhys.77.371>.
- [40] O.A. von Lilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, Int. J. Quant. Chem. 115 (16) (2015) 1084–1093, <http://dx.doi.org/10.1002/qua.24912>.
- [41] O.A. von Lilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, 2015. arXiv:1307.2918v4 [physics.chem-ph], <http://dx.doi.org/10.1002/qua.24912>.
- [42] S. De, A.P. Bartok, G. Csanyi, M. Ceriotti, Phys. Chem. Chem. Phys. 18 (2016) 13754–13769, <http://dx.doi.org/10.1039/C6CP00415F>.
- [43] M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Phys. Rev. Lett. 108 (2012) 058301, <http://dx.doi.org/10.1103/PhysRevLett.108.058301>. <http://link.aps.org/doi/10.1103/PhysRevLett.108.058301>.
- [44] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, J. Chem. Theory Comput. 9 (8) (2013) 3404–3419, <http://dx.doi.org/10.1021/ct400195d>.
- [45] J.M. Dieterich, B. Hartke, Mol. Phys. 108 (2010) 279, <http://dx.doi.org/10.1080/00268970903446756>.
- [46] J.D. Dieterich, B. Hartke, OGOLEM, A Framework For GA-Based Global Optimization. <<http://www.ogolem.org/>> (accessed: 2016-05-09).
- [47] N.O. Carstensen, J.M. Dieterich, B. Hartke, Phys. Chem. Chem. Phys. 13 (2011) 2903, <http://dx.doi.org/10.1039/c0cp01065k>.
- [48] J.M. Dieterich, U. Gerstel, J.-M. Schröder, B. Hartke, J. Mol. Mod. 17 (2011) 3195, <http://dx.doi.org/10.1007/s00894-011-0983-x>.
- [49] J.M. Dieterich, B. Hartke, Appl. Math. 3 (2012) 1552, <http://dx.doi.org/10.4236/am.2012.330215>.
- [50] Y. Li, B. Hartke, Chem. Phys. Chem. 14 (2013) 2678, <http://dx.doi.org/10.1002/cphc.201300323>.
- [51] U. Buck, C.C. Pradzynski, T. Zeuch, J.M. Dieterich, B. Hartke, Phys. Chem. Chem. Phys. 16 (2014) 6859, <http://dx.doi.org/10.1039/c3cp55185g>.
- [52] M. Dittner, J. Müller, H.M. Aktulga, B. Hartke, J. Comput. Chem. 36 (2015) 1550, <http://dx.doi.org/10.1002/jcc.23966>.
- [53] J.M. Dieterich, B. Hartke, Phys. Chem. Chem. Phys. 17 (2015) 11958, <http://dx.doi.org/10.1039/c5cp01910a>.
- [54] B. Bandow, B. Hartke, J. Phys. Chem. A 110 (2006) 5809, <http://dx.doi.org/10.1021/jp060512l>.
- [55] J.M. Dieterich, B. Hartke, J. Comput. Chem. 35 (2014) 1618, <http://dx.doi.org/10.1002/jcc.23669>.
- [56] B. Hartke, Chem. Phys. 346 (2008) 286, <http://dx.doi.org/10.1016/j.chemphys.2008.01.027>.

Publication: Optimization of Globally Optimal Catalysts

6.1 Scope of the Project

The following publication introduces most parts of the **GOCAT** design framework as main theme of this Thesis. As a proof-of-principle study it tackles a certain Menshutkin reaction and investigates many different settings for full-blown electrostatic global optimizations of catalytic effects. This ranges from very simple **GOCATs**, *e.g.*, just consisting of *one* partial charge, to successively more flexible **GOCATs** and consequently already portrays very well what, indeed, is possible by pure electrostatic catalysis. Admittedly, the chosen Menshutkin reaction shows a very clear tendentious effect of the catalysis with regard to the electric field strength and direction that can also be reproduced on a **DFT** level of theory and which is exactly the reason for addressing this reaction. Moreover, these **GOCAT** models are also contrasted with a common implicit solvent model (**COSMO**) in order to discuss the model restrictions. In this regard, critical evaluations of the models and the further future improvements are clearly addressed. The most self-evident one is the full *relaxation* of the **MEP** for the reaction *during* the global optimization to allow for **MEP** changes, from small ones to complete mechanistic alterations, by the catalytic surrounding. Hence in the following study, the so-called static or vertical **GOCAT** model is used, whereas the improvements are discussed in the next Chapter 7 in another context.¹

As Complementary Information for this publication in order to get the most out of this Menshutkin reaction, **EA** operator benchmarks that have been accomplished for this work are addressed afterwards (Section 6.3.1), some of the already used (and more) translation protocols between the levels of theory (Section 6.3.2) and finally a very clear-cut illustration of the aforementioned tendentious electrostatic catalysis effects (Section 6.3.3).

¹ Having already published this paper, the current author noticed (only due to an advice from a native English speaker) that “educt” is completely uncommon in English chemistry texts, contrary to the German usage. Hence, each “E” for “educt(s)” shall be changed to “R” for “reactant(s)” by the attentive reader in the following.


6.2 Publication Data and Reprint

<i>Reference:</i>	M. DITTNER, B. HARTKE, GLOBALLY OPTIMAL CATALYTIC FIELDS – INVERSE DESIGN OF ABSTRACT EMBEDDINGS FOR MAXIMUM REACTION RATE ACCELERATION, <i>J. Chem. Theory Comput.</i> 2018 , <i>14</i> , 3547–3564, DOI: 10.1021/acs.jctc.8b00151. ^[433]
<i>Submitted:</i>	February 10, 2018.
<i>Accepted:</i>	June 8, 2018.
<i>Contribution:</i>	Implementation of GOCAT as optimization class from scratch in OGOLEM with most of the features discussed in this Thesis, python scripts for automatic evaluations (ML), all calculations, analyses and discussions, mayor contribution to the text.
<i>Graphic:</i>	See p. 129 or the title page of this Thesis.
<i>ESI:</i>	Printed on pp. 299–362 in Appendix B.
<i>Copyright:</i>	Reproduced with permission from “Journal of Chemical Theory and Computation”. Copyright 2018 American Chemical Society.

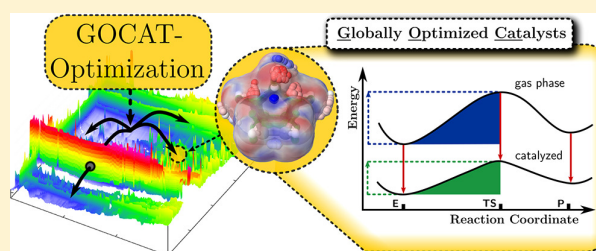
Globally Optimal Catalytic Fields – Inverse Design of Abstract Embeddings for Maximum Reaction Rate Acceleration

Mark Dittner[Ⓜ] and Bernd Hartke^{*Ⓜ}

Institute for Physical Chemistry, Christian-Albrechts-University Kiel, 24098 Kiel, Germany

 Supporting Information

ABSTRACT: The search for, and understanding of, good catalysts for chemical reactions is a central issue for chemists. Here, we present first steps toward developing a general computational framework to better support this task. This framework combines efficient, unbiased global optimization techniques with an abstract representation of the catalytic environment, to shrink the search space. To analyze the resulting catalytic embeddings, we employ dimensionality reduction and clustering techniques. This not only provides an inverse design approach to new catalytic embeddings but also illuminates the actual interactions behind catalytic effects. All this is illustrated here with a strictly electrostatic model for the environment and with two versions of a selected example reaction. We close with detailed discussions of future improvements of our framework.



1. INTRODUCTION

1.1. Inverse Molecular Design. Computational molecular design usually implies an *inverse* approach, i.e., starting from the desired properties one wants to reach and inferring possible molecular systems that could realize them. However, in chemistry and elsewhere, inverse problems are hard or impossible to solve in practice. A more feasible strategy is *forward* sampling: Starting from many “sensibly” chosen and chemically meaningful systems, the property in question is calculated *directly*—which now is a task in the usual quantum-chemistry domain of solving the Schrödinger equation—and then the system coming closest to the desired property value in question is selected, of course, via an appropriate algorithm and under sufficient automation (see refs 1 and 2 and references therein). Naive realizations of this forward strategy, for instance deterministic exhaustive enumeration or randomization without any chemical (meta-)heuristics, may hope for serendipity but will be highly inefficient: Chemical compound space^{3,4} is astronomically huge even for small compounds. Hence, a central task of molecular design is to strongly cut down on this vast search space without sacrificing all chances for finding good solutions. In the literature, many different approaches to this problem have been tried, of which we mention some typical examples: Obviously, expert chemical insight is useful,⁵ possibly in combination with simple molecular-orbital models^{6,7} or targeted quantum-chemical calculations.⁸ Another possibility is iterative buildup of complex target structures, with selection steps at each stage.⁹ Both deterministic¹⁰ and nondeterministic^{1,11,12} searches have been used to navigate the chemical space of real molecules. Additionally, virtual connections between real molecules have been exploited as shortcuts, in the form of alchemical derivatives and similar methods.^{13,14}

Also, data-driven machine-learning (ML)—reviewed also in refs 15 and 16—has been used to discover connections between structures and desired properties.^{17,18} Finally, recent applications of (inverse) optimization or ML-based inference tackling more complex properties of molecules also optimized heterogeneous catalysis models and subsequently their macroscopic properties,¹⁹ general (experimental) chemical reaction conditions,²⁰ and other properties (e.g., for drug discovery or solar cells).²¹

The above can be used to design molecular systems with essentially any desired property. One particularly interesting property, which we also address in the present work, is being an (optimal) catalyst for a given reaction. This has been characterized as one of the “holy grails of chemistry”.²² The most prominent examples of catalysts are of course not only enzymes²³ but also transition-metal complexes^{9,24} as well as heterogeneous²⁵ catalysts.

1.2. Related Work. Hence, our design target in the present work is catalysis. Of course, it is impossible to provide a comprehensive literature overview on catalysis here, even if narrowed down on analyzing and understanding catalysis mechanisms. Therefore, we only discuss a few exemplary studies here which focused on an *inverse approach* to actual catalyst design and with sufficient similarity to our own work.

Houk et al.^{26,27} introduced “theozymes”, as idealized enzyme active sites represented by a few concrete amino acid side chains, structurally optimized on the computer to achieve optimal transition-state stabilization for the desired target reaction (in a one-point/frame energy scheme, i.e., ignoring gradient

Received: February 10, 2018

Published: June 8, 2018

information and all other points on the reaction path). Subsequently, this was extended to the design of complete proteins (enzymes) by Mayo²⁸ and Baker²⁹ and is also often followed by experimental optimizations afterward.³⁰

Initially focusing on abstract electrostatic surroundings, Sokalski introduced Optimal Catalytic Fields (OCF)^{31,32} and his general Differential Transition State Stabilization (DTSS) concept,³³ both operating in a “differential” scheme (i.e., using a two-point/frame difference method). If electrostatic dominance is found, which is true in many cases, an optimal electrostatic scalar and vector difference field can be constructed: compare ref 34 for a concise description of the concept and references therein for similar electrostatics-based approaches,^{35–37} or for a newer application of the differential energetics idea see ref 38. In later work, Sokalski’s abstract OCF and DTSS procedures were linked with concrete active sites of enzymes.^{34,39–41}

Quite recently, Head-Gordon et al. also stressed the importance of the electrostatics for enzyme catalysis and explicitly developed and used MD-sampled averaged (projected) electrostatic fields for the essential breakage and formation of bonds during the reaction.⁴² Moreover, these electrostatic fields were then also used as guidance for (in silico) mutations of de novo designed enzymes mainly for electrostatic transition-state stabilization.⁴³

Another line of related work is “Gradient-driven Molecule Construction”,^{1,9,44} aiming at stabilization of otherwise nonstationary points on the potential energy surface (PES). Starting with a general description of an additional energy and interaction term in the Hamiltonian between a fragment and its embedding (“jacket” potential), this work investigated maximal (gradient-based) stabilizations, using different representations: 1.) one explicit partial point charge with continuously varying position, 2.) multiple point charges on prefixed positions (motivated by the transition-metal complex they were studying), or 3.) an even more fine-grained representation using the additional potential terms on the exchange-correlation functional integration grid.⁴⁴ This approach was recently extended to a *greedy* shell-wise construction framework.⁹ Note that in the present work we employ similar gradient criteria, as one of several fitness ingredients.

1.3. Catalytic Field Model. In our approach to catalyst design presented here, we reuse some conceptual ideas from the literature (mentioned in the previous subsection), but in reshaped form, as parts of a novel algorithmic strategy. Briefly described, for a single-step reaction to be catalyzed, we construct an abstract surrounding that maximizes an expected catalytic speed-up. This optimization is done with highly efficient nondeterministic global search tools. The final step of translating this abstract but optimized catalytic surrounding into a concrete molecular realization will be addressed in future work; here we focus exclusively on the initial optimization step.

There are no restrictions on the given reaction to be catalyzed. In the remainder of this Article, we will present results for a particular example reaction, but our overall concept is general and should work similarly for any other reaction.

On a boundary layer around this reaction center, enclosing the whole reaction coordinate (from reactants via the transition state (TS) to products), we then introduce what we dub “Globally Optimal Catalysts” (GOCAT in the following). Similar to Sokalski’s catalytic field, this is an abstract representation of a catalyst, but in contrast to Sokalski’s original idea, we understand this GOCAT as a more general, more complex entity that is to be optimized. In an advanced form, it

may include the following ingredients, in varying numbers, spatial positions, and strengths:

- partial (e.g., point, or multipolar) charges,
- van der Waals interaction centers,
- H-bonding centers (up to also other centers for, e.g., halogen bonding),
- others ...

For the present, first proof-of-principle application, we limit ourselves to partial charges. Note that the GOCAT concept is sufficiently extensible to also allow for a transition from the abstract to the concrete: To investigate specific effects in the context of enzymes, more concrete ingredients, for instance capped amino acids, could also be included. This would move the GOCAT concept more toward Houk’s theozymes.

This GOCAT abstraction layer simplifies and shrinks the search space so much that finding an optimal GOCAT can be addressed by a nondeterministic global search. For this purpose, we employ Evolutionary Algorithms (EA),^{45–47} because these can be easily mixed with other search paradigms without breaking the overall EA framework, e.g., with local search,^{48,49} with secondary order parameters (i.e., niching),^{48,50,51} or with locally focused exploration.^{52,53} This allows for high gains in search efficiency, by tuning the EA to the problem type at hand, as we have demonstrated by applying EAs to problems ranging from abstract benchmarks⁵⁴ via cluster structures^{50,55} and molecular design¹² to parameter optimization in reactive force fields.^{56–58}

Additionally, in the EA search for optimal GOCATs, we employ rather elaborate objective functions: They do not only focus on lowering the TS energy and on ensuring practically sufficient reactant affinity and product release but also include gradient data (not just energies), discretized (in so-called frames) along the whole reaction path. In part, this is necessary to avoid artifacts (lowering the energy at frame N could induce an energy rise at frame $N + M$ that annihilates any catalytic effect), and in part this allows design toward practically useful catalysts (reactant affinity, product release).

Of course, the GOCAT optimization by EA has to use some concrete energy/gradient calculational backend. At this point, our concept is again very general: On a force-field level, a realization would definitely be possible, e.g. employing partial point charges or multipoles as electrostatic GOCAT centers or tunable X-H dummy molecules as H-bond acceptors/donors. However, also a quantum-chemical realization is viable, on any theoretical level (semiempirical, DFT, ab initio): The electrostatic GOCAT centers can be conveniently modeled via the QM/MM approach,^{59,60} van der Waals centers by tunable rare-gas atoms, etc. [Of course, everything boils down to pure Coulomb interaction at the end,^{61–63} but in the usual interpretation of chemists, one could argue even about separate vdW interactions, H-/halogen-bond interactions, etc., which at least in principle could be represented by (fractional) GOCAT entities.] See Section 4 for further discussions about the actual model approximations resulting from that. For the present proof-of-principle demonstration, we have opted for a semiempirical level of theory, since this gives us raw speed (which is needed for global search in general and for exploring various conceptual alternatives) while retaining true quantum-chemistry ingredients, indicating that a transfer to higher-level methods is indeed possible.

As we will show below, GOCATs are complex and can exist in many near-equivalent forms. Therefore, we had to employ “big

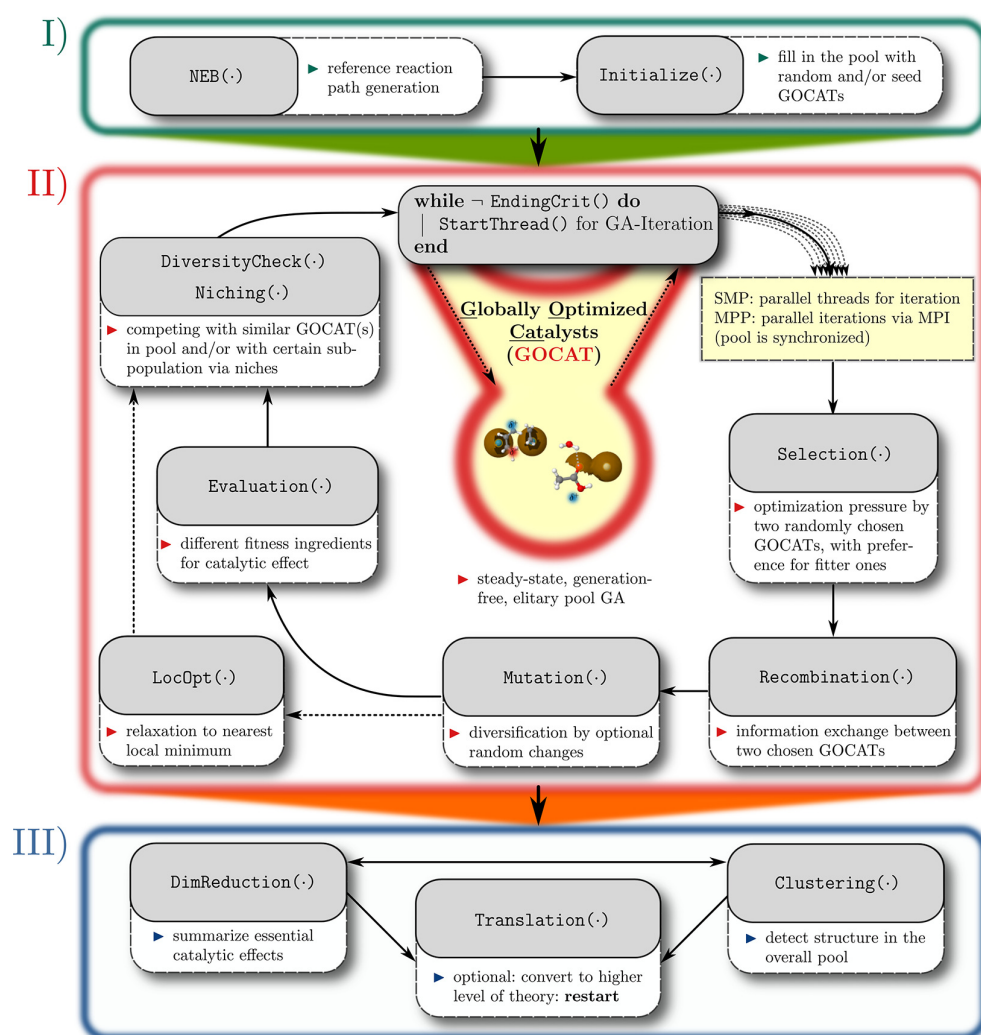


Figure 1. Overview of the procedure, including I) reference reaction path creation and initialization, II) GA optimization, and III) postanalysis via unsupervised learning/statistics and optional translation to the other level of theory, in which case I) restarts again. All essential algorithm blocks are described in the main text.

data” techniques to reduce this complexity toward results that can be interpreted and understood. We will briefly introduce only those techniques we actually need, skipping over the theoretical background and the related literature.

1.4. Disclaimers. Before going into the technical details, a few caveats are in order:

As indicated, we are limiting ourselves here to abstract GOCAT optimization, leaving the necessary translation of such a GOCAT to a real-world catalytic molecular frame for future work. Nevertheless, after complexity-reduction, already such an abstract GOCAT does provide insights into how and why catalytic effects arise.

We are also limiting ourselves to electrostatics (point-charge GOCAT centers) here. For several years, there has been a fierce debate in the literature if this electrostatic effect indeed is the *only* catalytic effect, as forcefully argued by a recent Nobel laureate, Arieh Warshel, in a series of papers,^{64–66} while several other authors equally forcefully insist on the existence of further catalytic effects, e.g., based on enzyme dynamics or a broad range of not purely electrostatic enzyme–substrate interactions.^{67–71} We do *not* take sides in this debate: The possibility to argue for

“electrostatics only” supports our choice to initially focus only on this ingredient, but the GOCAT concept is broad enough to include other effects.

Homogeneous catalysis frequently involves transition-metal (TM) atoms as crucial ingredients. In later stages of GOCAT development, TM centers may also enter our catalyst design strategy, but we deemed this too complex as a first step.

Last but not least, many real-world catalytic processes involve many elementary reaction steps, frequently arranged in a cyclic fashion and with protons and/or electrons being shifted from one molecular entity to another. Addressing or setting up such multistep catalytic cycles is *not* yet our aim here; instead we merely try to arrive at significantly reduced energy barriers for one reaction step only (even explicitly disfavoring the appearance of intermediate minima). Obviously, one-step catalyst design can be combined with emerging methods for reaction network exploration^{72–77} to arrive at strategies generating and optimizing multistep catalytic cycles. Some of these points will also be discussed in Section 4 again.

The remainder of this article is organized as follows: The EA details including the actual objective function, niching, and post-

EA clustering procedure are given in Section 2. In the beginning of Section 3, our example reaction is illustrated. Subsequently, results for GOCATs of increasing complexity are illustrated together with discussions of the observed catalytic effects. We will then continue with a further (meta-)discussion of the used GOCAT in Section 4 and end with a conclusion in Section 5.

2. IMPLEMENTATION AND METHODS

In this first GOCAT illustration we used a predefined reaction path. Its structural coordinates remain the same throughout the GOCAT optimization. This path is discretized into a finite number of “frames”. The central GOCAT optimization aim is to change the energies at all of these frames in order to accelerate the reaction, within further restrictions explained below. The GOCAT consists of partial charges placed on a van der Waals surface. This surface encloses all frames along the whole reaction path, and the partial charge values and positions are not allowed to vary from frame to frame.

Thus, we have a mixed optimization problem containing two different groups of parameters: Cartesian coordinates on a curved 2D surface and partial charge values. As anticipated (and shown later on), this optimization problem needs an efficient global optimization algorithm having to treat multiple different local minima in an unbiased way. For this purpose, we employed and extended our global optimization framework OGOLEM.⁴⁹ [At the time of writing, the current work as well as the extensions that are described in ref 57 is an own *local* fork of the OGOLEM project. The latter can be found at ref 78; for the former contact the authors.] The core of this framework is a generation-free⁷⁹ Genetic Algorithm (GA), including options for problem-specific “phenotype” operators,^{50,52,53} niching,^{48,51} and further ingredients that increase optimization efficiency. Additionally, OGOLEM features advanced parallelization techniques that allow for failsafe computing distributed across strongly heterogeneous hardware⁸⁰ and for highly adaptive parallel runs that exploit traditionally inaccessible scheduling gaps.⁸¹ Partially, existing OGOLEM algorithm concepts have been reused or adapted to the current problem. These newly implemented GA details will be briefly described in the following. For general descriptions of our GA framework and its more standard ingredients, we point to former work, see refs 49, 51, 57, and 80.

In Figure 1, the general procedure is sketched. In the following, all the different steps are described, in the order of their importance.

2.1. General Optimization Steps and Quantum-Chemical Theory Levels. For the present GOCAT optimizations, the following key steps were executed; further explanations are given below (in the following, educt is abbreviated as E, transition state is abbreviated as TS, and product is abbreviated as P):

- Part I), NEB (\cdot): Without any GOCAT, the reaction path itself is optimized, at the level of theory used later on. About 20 frames between E and P are defined and constrained during the GA, as well as a common van der Waals (vdW) surface enclosing all these 20 frames. Initialize (\cdot): Usually a population (of about $n = 600$ GOCATs) was generated, where the random distribution of charges was weighted by the exposed vdW surface. Each GOCAT consists of a fixed number N_{Ch} of partial charges constrained to be positioned on the vdW surface and constrained to charge values between preset minima/maxima throughout.

- Part II): Global GA optimization of the GOCAT population: The genotype dimension is about $4 \cdot N_{\text{Ch}}$ (one charge value and three Cartesian coordinates for each charge). About 10 separate GA runs each for a separate population were executed. LocOpt (\cdot): The final solutions of all runs were *locally* optimized (i.e., in the last GA-iteration for each GOCAT in part II)) with respect to the same objective function as in Evaluation (\cdot) (vide infra), using the gradient-free BOBYQA⁸² optimization algorithm in our framework, and then accumulated into a common database.
- Part III): This database (of a given N_{Ch}) of all solutions was analyzed, using mainly cluster analysis methods for identifying meaningful patterns/domains of dominant charge embeddings.

The main observations and discussions presented below will focus on the results *up to and including this last point*.

Our GOCAT framework presented here is intrinsically independent of the level of theory at which the actual calculations are performed. For efficient exploration of the GOCAT search space and to allow a sufficiently wide testing of various changes in algorithm ingredients, we decided to employ a low level of theory in this first GOCAT study. Hence, for all the above steps, we used MOPAC⁸³ with the PM7⁸⁴ Hamiltonian and parameters and with the (classical) QM/MM coupling scheme available there.⁸⁵ For the exploratory calculations at higher levels of theory mentioned in the next paragraph, we used the ORCA⁸⁶ program suite and mainly DFT calculations with the PBE0⁸⁷ functional and def2-TZVP⁸⁸ basis set.

To explore possible contacts to higher levels of theory, we also added the following steps corresponding to part III) and specifically Translation (\cdot), leading to a restart beginning at I) and including the older results as starting GOCATs:

- For several of the most important GOCAT clusters from the cluster analysis, both the best ranks and the nearest neighbors to the (artificial) cluster centers were collected for creating a most diverse input guess for the next steps (seeds).
- To allow for structural changes between the different levels of theory, the vdW surface was redefined for the higher level of theory, and the GOCATs were mapped to this new vdW surface.
- Another GA-based global optimization on this higher level was executed.
- LocOpt (\cdot) based on energy, not on the objective function: Additionally, within their given charge embeddings, the reaction frames of the final GOCATs were then locally optimized once more, not with respect to the GOCAT-search fitness function, but structurally relaxed to locally minimal energies or to reach the next first-order saddle point. This caused them not only to leave the vdW surface and the prefixed reaction path slightly but also to reach converged new stationary points on the new level of theory (with final additional frequency checks to verify the stationary points E, TS, and P).

See sections Section 4.1 for further discussions of these higher-level calculations.

2.2. Objective Function (Evaluation (\cdot)). All properties to be optimized were accumulated in one aggregate sum as an objective function that is to be *minimized* via GA and will also be called fitness function synonymously. After extensive testing, this included finally the following:

1. $\Delta E^\ddagger = E_{\text{TS}} - E_{\text{E}}$, the (pure electronic) energy barrier between the TS and E should be minimal.
2. The TS should at least be *stabilized* with respect to the reference path without any GOCAT.
3. No new intermediate minima on the reaction paths are allowed, i.e., the reaction profile should remain unimodal.
4. The minima/maximum energy frames (E, TS, P) should not change their positions in coordinate space too much: A fitness penalty was added if an extremum of the GOCAT energy profile moves more than 2 frames off the reference case. However, this penalty does not come into play very often.
5. The gradient norms of the E, TS, and P frames should at least be lower than a threshold, i.e., they should approximately retain their character as stationary points.

Note that the gradient restraints in item 5. are more important than they may seem at first. Without such gradient restraints, E, TS, and P could be arbitrary points, not necessarily stationary ones. Hence, as a limiting case, the reaction profile could be transformed into any arbitrary path on the GOCAT-modified PES, even into an equipotential contour line, with no barrier at all but also without the defining minimum-energy pathway characteristic (the PES-gradient component perpendicular to the reaction coordinate is not (close to) zero anymore). This could be dubbed overfitting, since the other requirements item 1.–item 4. are fulfilled perfectly, but such a solution would not be a reaction path anymore.

Similarly, item 3. is vital to avoid artifacts. During the GOCAT optimization, each reaction path of a GOCAT solution was analyzed by reading out all maxima and minima and calculating artificial barriers between those, which then enter into the overall fitness. This is necessary as otherwise a new minimum could occur, for instance between E and TS. As a result of that, the effective barrier of a new maximum energy frame minus a minimum energy frame could be higher than the reference barrier. Hence, item 3. effectively enforces a unimodal reaction profile, increasing from E to TS and then decreasing again, such that the final and only barrier included in the globally optimized GOCATs is the one between TS and E. [Otherwise arbitrary intermediate minima and maxima in the reaction profile could result, turning the single-step reaction into a multistep one, not on purpose but merely as an overfitting artifact. In other systems studied so far, with a nonlinear and longer reaction coordinate (hence also containing more frames), this item becomes even more important. So, charting and relating multiple such frames (not just 2 frames) is necessary and is done via this item.] Note also, that no stabilization or destabilization (as another possible root of barrier decrease) of the educt frame is enforced in the present objective function definition.

Different weights on the above ingredients were tested, as well as different restraint functions for separate terms. For the results of this paper, the final settings were the following: The barrier, item 1., is highly weighted, since this can be expected to be the main property for creating a catalytic effect. If the TS is not stabilized corresponding to $\Delta E_{\text{TS,stab.}} > 0$, item 2., a penalty of $f(\Delta E_{\text{TS,stab.}}) = p \cdot \Delta E_{\text{TS,stab.}}^2$ was added, with $\Delta E_{\text{TS,stab.}} = E_{\text{TS,GOCAT}} - E_{\text{TS,ref.}}$ and a given weight p . Such a quadratic penalty was also used for all three gradients, item 5., for all $\|\nabla E_{\{E, \text{TS}, \text{P}\}}\| > 10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ (explicit weights of the objective function terms are given in the SI).

The only two nonzero terms within the objective function for the globally and locally optimized solutions shown in this work

usually refer to the barrier, item 1., between E and TS and to the gradient norms, item 5., of the stationary points. All other (penalty) terms, including also the shifted frames, item 4., are already zero after complete optimization. In other words, those latter ingredients help to guide the initial walk through the search space, immediately penalizing nonchemical GOCATs, while a "trade-off" between the former ingredients dominates the endgame and the final fitness ranking.

As in another OGOLEM application,⁵⁷ this insight was exploited to reduce computational expense: In the fitness calculation, the barrier item 1. and gradient item 5. items were calculated before the reaction profile analysis item 3. If the fitness accumulated from the barrier item 1. and gradient item 5. items was already worse (higher) than the worst current GOCAT of the GA-population, remaining objective function terms (reaction profile analysis item 3. and the others) were completely neglected and the fitness calculation was stopped, without any influence on our GA progression. Thus, this "Immediate Fallback"⁵⁷ was used as a search space reduction technique, saving many QM/MM single-point calculations especially at the end of the GA when the population is already almost converged.

2.3. GA Setting. For the GA, we used a mix of different operators: *Recombination*(\cdot): 1.) A genotype crossover, exchanging both 3D Cartesian information and the charge values between the parents, cutting the genotypes as string of 4D charge entities at several places, and 2.) a phenotype operator (similar to the typical phenotype cut-and-splice operator⁸⁹), cutting at 3D planes through the GOCAT and splicing together different parts. Note that all essential geometric movements by the operators are mapped onto the common vdW surface after each operation again. *Mutation*(\cdot) and *random Initialize*(\cdot): As mutation, we used also a mix of different operators: 1.) with small probability, a nullary random reinitialization: A packing operator directly on the vdW surface of the reaction, which simply is the random initialization operator for generating the GOCATs in the starting pool. 2.) Usual unary mutation operators for random Cartesian displacements and charge values of a subset of the charges around the current values (within a Gaussian distribution) of the parents. This choice of operators is the result of an initial benchmarking to increase the general GA efficiency for this kind of optimization, i.e., a greedy brute-force meta-optimization of those GA settings was done beforehand but will not be detailed here.

Niching(\cdot): Moreover, we utilized our niching framework⁴⁸ in order to conserve diversity within the population during the GA and to decelerate premature convergence. To this end, a niching version was implemented that is similar to the "dynamic" niching based on the Coulomb matrix which was explained in detail elsewhere.⁵¹ In the present case, the electrostatic potential (ESP), φ_{ESP} , introduced by the charges of each GOCAT serves as molecular descriptor: The net effect (superposition) of any structural composition of partial charges around the reaction path was calculated at the atoms of selected frames of the path (E, TS, P) as

$$\varphi_{\text{ESP},j} = \sum_i \frac{q_i}{r_{i,j}} \quad (1)$$

(in atomic units) where i counts the partial charges with values q_i and J counts all separate atoms of the core-frames used with distances $r_{i,j} = |\mathbf{R}_i - \mathbf{R}_j|$ between their position vectors $\mathbf{R}_{\{i,j\}}$. (Incidentally, this is also the combined quantity of the MM

atoms (charges) added to the Hamiltonian of the QM part in the QM/MM coupling scheme used in the semiempirical level of theory.⁸⁵) To define the niches, different procedures like a minimal threshold on a norm of the difference vector between two such ESP-vectors of two individuals were used as well as a vector-element-wise comparison of two individuals with a minimum number of almost identical ESP values. If such defined similarities are less than a threshold, the individuals were binned to the same niche. As only a maximum number of three individuals per niche were allowed during the GA, new individuals that are supposed to be similar to the ones already in the population just compete with the individuals of that niche as subpopulation instead of competing with the complete population. (Thus, at least ≥ 200 niches within the populations of $n = 600$ were always conserved in this concrete case.) Note that this scheme is similar, but not identical, to a real full-blown cluster analysis based on agglomerative hierarchical clustering described below and used as a postanalysis method. [One could think of this scheme as a greedy, linear-scaling version of single-linkage hierarchical clustering. So, small predefined niche sizes are meaningful because of the *chaining phenomenon* otherwise: Just *two* close individuals would lead to a merge of two niches, without enforcing “compactness” within the whole niche as no other maybe more distant individuals of that niche are considered for the assignment.⁹⁰ We favored this on-the-fly clustering scheme in order not to introduce new serial bottlenecks into our highly efficient generation-free pool GA.⁷⁹ Yet, in a quite similar context, full-blown cluster analysis at a certain frequency was recently also employed by others.⁹¹]

Selection(\cdot): A rank-based parent selector was applied with a Gaussian distribution centered at the lowest ranks (i.e., the best current individuals), for choosing one parent with a higher probability at the better end of the pool. The other parent was chosen with a uniform distribution. This is standard practice to support both exploration and exploitation.

2.4. Reaction Path. **NEB(\cdot):** For the reaction path optimizations we used mainly a nudged elastic band (NEB) optimization (with a climbing image),^{92,93} newly implemented in OGOLEM (and inspired also by the implementation in ASE⁹⁴), together with the FIRE⁹⁵ local optimization algorithm. To our experience, this local optimization algorithm leads to a more fail-proof NEB optimization than other local optimization algorithms, because of the nonconservative NEB forces.⁹⁶

Note that of course any scheme for the computation of a minimum energy path (MEP) could have been used as a prestep to the GOCAT global optimization. But for later (future) flexible reaction paths that are locally relaxed during the GA (cf. the outlook in Section 4.2), NEB will come in handy again (i.e., NEB(\cdot) would be part of Evaluation(\cdot) itself). Also, we completely aligned each frame of the NEB to each neighbor frame(s) before NEB-distance calculations, essentially removing external degrees of freedom (overall translation and rotation) which was now also published by others elsewhere.⁹⁷

2.5. Cluster Analysis. In the domain of Machine Learning (ML) and cheminformatics in general, a great deal has already been said about molecular descriptors and similarity metrics between chemical entities (see e.g. refs 98–101). In this specific case, the 2-norm of the difference vector of the ESP vectors of eq 1, φ , was used as a similarity measure, $d(\varphi_n, \varphi_m)$, between two individuals, n and m

$$\|d(\varphi_n, \varphi_m)\|_2 = \left(\sum_j (\varphi_n - \varphi_m)^2 \right)^{1/2} \quad (2)$$

In the case of PM7 level of theory calculations, the reaction path was of C_{3v} symmetry, which was also taken into account during similarity calculations.

Moreover, for cases of almost identical electrostatic embeddings, measured via differences of φ (eq 2), but different structural patterns, the so-called Bag of Bonds (BoB) descriptor was applied (in a slightly adapted version).¹⁰² In short: Based on the Coulomb matrix descriptor,¹⁰³ a vector of concatenated bags, each representing a certain type of “bond”, is formed

$$\frac{q_i q_j}{|\mathbf{R}_i - \mathbf{R}_j|} \quad (3)$$

i.e., essentially their attractive or repulsive Coulomb interaction potential energy, but this time including also the partial charge entities (GOCAT). These bags of such terms of eq 3 for qualitatively different bonds or interactions types ($\{C-C, C-N, Ch-C, Ch-Ch, \dots\}$, with Ch as “charge”), are each sorted separately in order to get a permutationally invariant representation (besides the overall translational and rotational invariance as redundant internal coordinate representation) and concatenated together to one (super)vector for each GOCAT. Note that in the original formulation,¹⁰² the positive nuclear charges are used, whereas in this electrostatic GOCAT embedding, the core-frames used (E, TS, P) are constrained. So, their atom-pairwise distances as well as their “identities” never change, because in the present case no similarity comparisons of entities throughout the chemical compound space take place like in other ML-approaches for learning properties of *qualitatively different* molecules or molecule classes. Thus, just the partial charges of the charge entities themselves are used in the numerator, while the core atom charges are neglected. Then, in most cases also the Euclidean norm of the difference of two BoB representations for two individuals was computed.

Both descriptors were used for identity checks (filter procedure): If ESP difference vectors (eq 2) between two individuals in the accumulated database are similar (lower than a threshold of usually $1.0 \text{ kcal mol}^{-1} \text{ e}^{-1}$), the worse individual of such a comparison is erased from the database, if they are also similar with respect to their BoB representation, to test for overall structural (dis-)similarity (compare with ref 91, where a similar identity check approach was described).

Clustering(\cdot): For the post-GA analysis part (part III)), we applied the agglomerative hierarchical cluster analysis as an unsupervised Machine Learning method implemented in SCIPY¹⁰⁴ and SCIKIT-LEARN.¹⁰⁵ The general procedure thereby was as follows:

- Each accumulated database for a given GOCAT size (about 6000 individuals) is filtered with respect to fitness and ΔE^\ddagger .
- Filter out further GOCATs by erasing (almost) identical ones (identity check): After local optimization and accumulation, identical solutions might be present as the niching erased identical ones only within the *separate* runs. Thus, a filtered database of about ~ 5000 resulted.
- Full hierarchical clustering using the ESP difference via eq 2 as a similarity measure was used. Usually, for our purpose, the *average linkage* strategy (unweighted pair

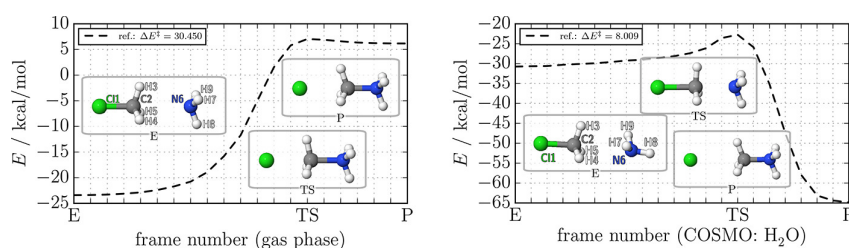


Figure 2. PM7: reference gas phase (left) and reference COSMO (H_2O) (right) NEB reaction path, given per frame. Insets show the 3 stationary points. Additional details are compiled in Table 1.

group method with averaging, UPGMA) was the best to identify main domains of electrostatic embedding without deforming the clustering space and with the best correlation between composed cluster similarities and the original similarities between the individuals themselves.

- As our main two targets here with this analysis are¹⁰⁶ 1.) to gain some insight into the underlying structure (hypothesis building) and especially 2.) to compress the data for summarization or organization of the results, we used different cutting schemes of the dendrogram (i.e., final number of clusters), at the end validated mainly through visual inspection of the results.

`DimReduction(·)`: Additionally, we used standard lower-dimensional projection techniques like Multidimensional Scaling (MDS) and Principal Component Analysis (PCA) to get further insights into the structure of the database.⁹⁰

3. RESULTS AND DISCUSSION

As an example reaction, we investigate the prototype Menshutkin reaction¹⁰⁷ $\text{H}_3\text{CCl} + \text{H}_3\text{N} \rightarrow \text{Cl}^- + \text{H}_3\text{NCH}_3^+$ as an $\text{S}_{\text{N}}2$ reaction with uncharged educt molecules and products with a net charge separation. This reaction usually is endothermic with a high barrier in the gas phase. With increasing polarity of a surrounding environment (solvent), the reaction is more favorable both thermodynamically as well as kinetically due to the stabilization of the charged products and the partial charge separation at the transition state (for some further energetics of this reaction and the impact of solvation and other surroundings, we point to refs 108–110 and references therein).

Before we are composing partial charges around the complete reaction coordinate, first of all the (precalculated) reference paths are given in Figure 2. Additional geometric measures are specified in Table 1. For the remainder, the 3 explicitly given frames of the stationary points (given in the insets) are superposed in all of the GOCAT pictures that follow. Here, we will delay the discussion of the details of the reaction path to a later Section, where the geometric differences (and with this the possible Cartesian placements of the embedding and corresponding electrostatic influence) are scrutinized.

3.1. Simplest Case: $N_{\text{Ch}} = 1$. In the following, we use the nomenclature $r_n\text{-}c_m\text{-}n_x$ as the full cluster name for the cluster analysis part: r_n defines the best rank or position of each separate cluster in the complete database, with lower numbers for n (starting at zero) corresponding to lower (= better) fitness/objective function values. c_m is a numbered nickname for the cluster itself, and n_x denotes the cluster size, i.e., the number of individuals in that cluster.

Table 1. PM7: Geometric Measures (Bond Distances and Bond Angles) of Given Frames in the Insets of Figure 2^e

level of theory	frame ^a	$r_{\text{CN}}/\text{\AA}$	$r_{\text{CCl}}/\text{\AA}$	$\angle(\text{HCN})/\text{deg}$	$\angle(\text{HCCl})/\text{deg}$
PM7 gas phase ^b	E	3.031	1.775	71.5	108.5
	TS	1.688	2.341	101.5	78.5
	P	1.543	2.589	109.5	70.5
PM7 COSMO: H_2O ^d	E ^c	3.123	1.783	72.0	108.0
	TS	2.146	2.035	84.3	95.7
	P	1.500	3.189	110.6	69.3

^aMore precisely: E will be the reaction complex (RC), and P will be the product complex (PC) below within the surrounding GOCAT, see Section 4.2. ^bLength of total reaction coordinate: 1.963 Å, TS at 1.737 Å. ^cLength of total reaction coordinate: 3.307 Å, TS at 2.021 Å. ^d NH_3 is rotated off the central attack line. ^eAtom numbering suppressed here (Cl1, C2, N6), while H3–5 sits at C, H7–9 at N (one H atom is used in the table because of symmetry).

For getting an impression of how the reaction can already be manipulated by adding just one single charge ($N_{\text{Ch}} = 1$) on the common vdW-surface, Figure 3 shows a dendrogram for hierarchical clustering of the corresponding database. There in the insets, all electrostatic GOCATs consisting of one partial

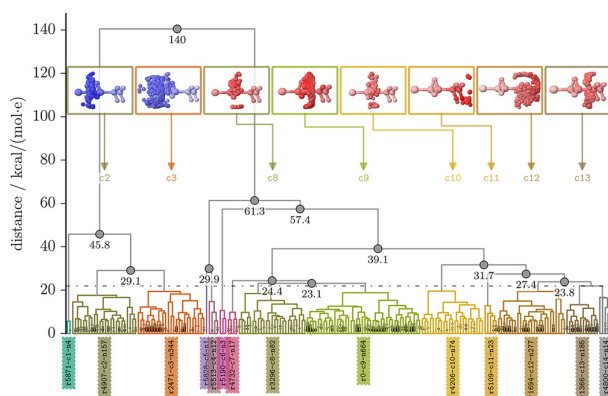


Figure 3. $N_{\text{Ch}} = 1$ GOCATs (PM7): Dendrogram of the final database with 1857 nonidentical individuals using the average linkage strategy. As distance metric, eq 2 based on eq 1 was used. For illustration purposes the dendrogram was cut to reach 14 different clusters, effectively cutting at a distance of (directly below) 21.89 kcal mol⁻¹ e⁻¹ (dotted line). In the insets, all corresponding overlain GOCATs are shown for some selected bigger clusters, colored between red and blue for negative and positive partial charges and ESP values. This dendrogram is truncated, i.e., the small ellipses at the leaves show additional branching points of the binary tree that are not plotted. The main branching points of the clusters are separately annotated by the numbered dots.

charge chunked together to clusters are overlain with also overlain selected frames of the reaction for E, TS, and P. Each partial charge and also the core Menshutkin reaction frames are colored between red and blue for negative and positive partial charges ($[-0.17, +0.17]$ e) and ESP values ($[-36.77, +36.77]$ kcal mol⁻¹ e⁻¹) on those reaction frame atoms, respectively. Because of the 6-fold symmetry at PM7 (C_{3v}), each mean ESP vector (for the core atoms) was calculated by starting at the best (lowest) rank of the cluster as a reference and mapping each residual GOCAT of the same cluster to that same symmetry representation before averaging each ESP value at the core atoms. Complementary to this, by showing a 2D MDS, Figure 4

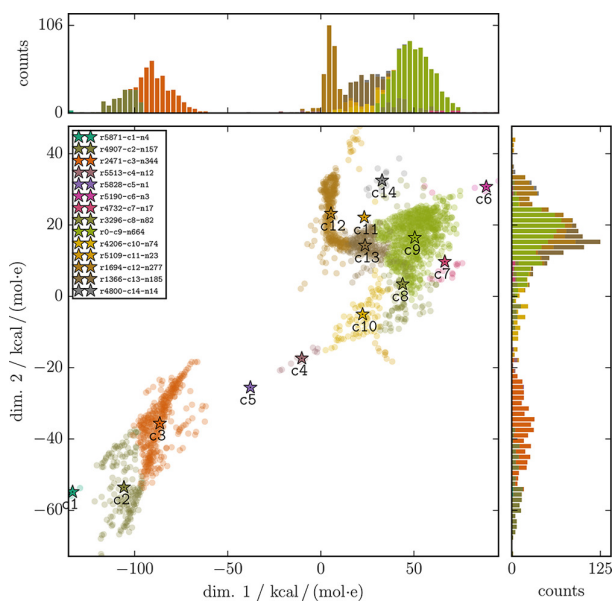


Figure 4. $N_{\text{Ch}} = 1$ GOCATs (PM7): MDS as 2D projection of the higher dimensional ESP-distance data is shown for illustrating the core cluster regions. Colored stars are the calculated mean 2D-coordinates (centroid) of a cluster. The (unnormalized) stacked histograms illustrate the number of individuals in those clusters.

illustrates the separability of our ESP space in 27 dimensions (3 · N_{atom} with N_{atom} atoms per frame and 3 frames used: E, TS, P). Individuals with bigger distances (for instance, the biggest distance between the lower left corner vs the upper right corner) will be mapped to qualitatively different electrostatic embeddings: So, one can divide the database roughly in either positively charged GOCATs or negatively charged ones. The former happen to lie at the lower left corner (big clusters: c2, c3), and the latter lie at the upper right corner (big clusters: c9, c12, c13) in this 2D projection. This also maps to the highest node (entry-point) of about 140 kcal mol⁻¹ e⁻¹ distance in the dendrogram (Figure 3). In this case, we deliberately stopped at 14 clusters during hierarchical clustering for illustration purposes: Some outliers (small clusters) define their own clusters, while the main regions (funnels) of the final fitness surface are described by a few clusters in this case.

Looking at the overlain GOCAT pictures in the dendrogram, this 2-fold division of the population leads to the following two funnels: On the one hand, the best catalytic effect (with mean barriers of the cluster of about $\Delta E^{\ddagger} = 25.4$ kcal mol⁻¹) corresponding to the lowest fitness function values, i.e., rank 0 (r0), is found in the main basin of attraction on the fitness

surface around cluster c9 (see the GOCAT inset pictures in Figure 3 or in Figure 5). As could be expected beforehand, a

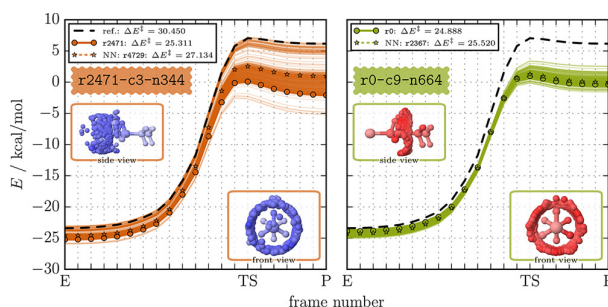


Figure 5. $N_{\text{Ch}} = 1$ GOCATs (PM7): (Preoptimized) reaction path for the PM7 reference calculation is shown as well as the path within the electrostatic GOCATs of the 2 main clusters (c3, c9). The lowest rank of each cluster (circles) and the nearest neighbor (NN) to the mean ESP vector of that cluster (stars) is plotted separately. All other individuals of the cluster are also plotted with thin lines to illustrate the spread. Two images in different perspectives of the overlain individuals are given in the insets.

negative charge (about -0.11 e) sitting between the methyl C atom and (attacking) N atom of ammonia is beneficial for the catalytic effect. This stabilizes especially the building of the positive charge (on C and N), resulting finally in the H_3NCH_3^+ product ion. Multiple similar and overlain GOCATs of this kind form a “ring-like structure” in the Cartesian domain. This corresponds to a robust solution with many qualitatively similar charges in that region.

On the other hand, one positive charge (about $+0.09$ e) in cluster c3 can be placed around the Cl atom, again especially with influence on the product side, when the Cl^- anion forms, resulting in mean barriers about $\Delta E^{\ddagger} = 26.8$ kcal mol⁻¹. A “fitness surface” showing these main basins of attraction (see Figures S1 and S2) and the corresponding statistics (Table S1) for this database as well as explicit single GOCATs are given in the SI.

The resulting reaction barriers within the electrostatics of the GOCAT are shown in Figure 5 together with the reference barrier in the gas phase. The best rank of the clusters as well as a representative for the cluster center is given explicitly, while the rest of the cluster is plotted with thin lines to show the spread and density of the clusters. Note: Each GOCAT in those plots leads to a final barrier that is less than the PM7-reference of $\Delta E^{\ddagger} = 30.45$ kcal mol⁻¹ corresponding to a catalytic effect, has gradient norms about 10 kcal mol⁻¹ Å⁻¹, and shows a smooth reaction path as defined in the objective function and enforced by the filtering processes of the database while having qualitatively different influences on the reaction (positively vs negatively charged at different places). In almost all cases, the stabilization energy (negative energies) relation $\Delta E_{\text{P,stab.}} < \Delta E_{\text{TS,stab.}} \ll \Delta E_{\text{E,stab.}}$ holds: The late TS is stabilized also severely but slightly less than P, whereas the neutral E is less influenced leading to an effective (electronic energy) barrier decrease. The remainder of these GOCAT-size reaction profiles of all the other 12 clusters is given in the SI for comparison.

3.2. More Charges: $N_{\text{Ch}} = 10$ (Non-Neutral Summed Charges). Going to multiple partial charges as abstract GOCATs increases the complexity of the optimization but also allows for creating electrostatic embeddings that address the core-atoms separately, providing some or all of them with

individually optimized surroundings and that may also produce correlated or synergistic effects. We have tried several different N_{Ch} values but present here only the $N_{\text{Ch}} = 10$ size. This is already a converged end point of a general trend, starting at $N_{\text{Ch}} = 1$, reaching $N_{\text{Ch}} = 3$ (SI), and not changing qualitatively at $N_{\text{Ch}} = 20$ (not shown). First, without any additional constraints we optimized GOCATs that showed up a very positive or negative overall summed charge. The observations we made can be summed up as follows:

- More nonredundant GOCATs are possible (bigger search space due to higher complexity/more freedom and accordingly the already bigger database shown will not be an exhaustive representation of the bigger search space in this case).
- GOCATs happen to lie on a straight line within dimensionality reduction, i.e., the database might even be reduced to *one* (essential) dimension (due to gauge freedom). This provides a hint that one specific qualitative solution (ESP relation on the atoms) is most important for the catalytic effect and that now embeddings with enough freedom or flexibility (of 10 charges) can discover this.
- Very broad ESP range, although the biggest clusters build up at mainly overall neutral GOCATs and correspond to broad basins of attraction (funnels) on the fitness surface.
- There is of course no bijectivity of Cartesian placement of partial charges and the resulting ESP at the core atoms anymore (which is also the reason for the additional BoB descriptor mentioned above in the filtering process). This results in different possible Cartesian domains on the vdW surface for similar GOCATs with regard to electrostatics.

A detailed discussion and illustration is given in the SI. To erase this artificial freedom, in the remainder of this article an additional constraint on overall charge neutrality during GOCAT optimization is added in.

3.3. $N_{\text{Ch}} = 10$ with Overall Charge Neutrality. To eliminate the redundant total GOCAT charge, we repeated the GOCAT optimization for $N_{\text{Ch}} = 10$ of the previous Section 3.2 but now with the additional constraint of overall charge neutrality. This indeed has the desired effect: The dominance of one single dimension in the MDS analysis disappears (SI: cf. Figure S22), revealing nontrivial structuring in the other GOCAT characteristics, see Figure 6, that will be discussed below. Similar plots for hierarchical clustering—from which one cluster is presented in the following—for this case are moved to the SI.

A stacked histogram of the ESP values of the by far biggest cluster ($r0\text{-}c11\text{-}n3131$) of the neutral case, which also includes the best rank found, is illustrated in Figure 7. There, the biggest cluster with about 3000 GOCATs can be distributed into 3 main domains (Gaussian-like distributions in the histogram). The ESP values at the Cl atom (green ones) almost stay the same upon moving from educt (E–Cl1) to product (P–Cl1)—which then has anionic character—with a positive $\varphi_{\text{ESP,Cl}} = 14$ kcal mol⁻¹ e⁻¹. Also the ESP values at the N atom (red) can be found at about $\varphi_{\text{ESP,N}} = -4$ to -7 kcal mol⁻¹ e⁻¹ from educt to product. The C atom is even more negatively embedded (blue) and shows a shift from $\varphi_{\text{ESP,C}} = -5$ to about -15 kcal mol⁻¹ e⁻¹ on average. This comparatively very large shift at the C atom, the small shift at the N atom, the absence of a significant shift at the Cl atom, and all average values match nicely with organic chemistry expectations of partial charge changes at these atoms,

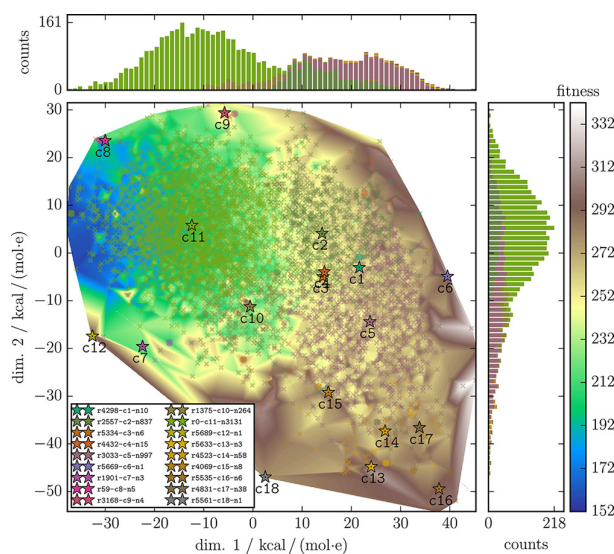


Figure 6. $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral): MDS 2D projection of the higher dimensional ESP-distance of a database of 5388 nonidentical GOCATs. Also, a linearly interpolated fitness surface (color map) is plotted. Besides the cluster means (stars) also the best rank of that cluster (the individual indicated by r_n in labels of the legend) is plotted as a circle. For the other illustration details, compare with Figure 4.

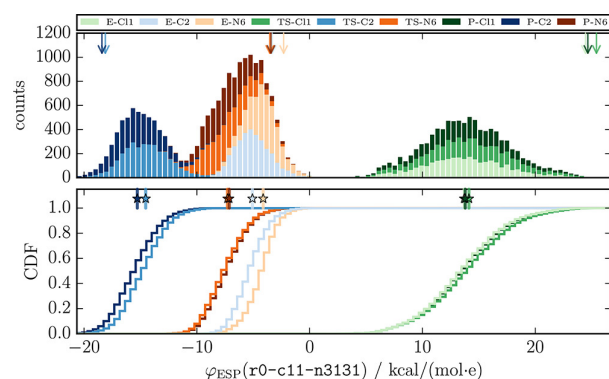


Figure 7. $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral): Top: Stacked histogram of all the ESP values at the Cl, C, and N atoms of the E, TS, and P frames of the biggest *neutral* cluster $c11$. Arrows indicate the explicit ESP values of the best rank within their distributions, r_0 (given in Figure 8), following the coloring of the 9 separate atoms shown in the legend. Below: Cumulative distribution function showing the spread or (if present) skewness; vertical bars with stars are the computed *average* ESP values of that cluster at the corresponding atoms. All values in kcal mol⁻¹ e⁻¹ for E, TS, and P frames (standard deviation in parentheses): Cl: 13.70(3.67), 14.12(3.86), 13.79(3.71); C: $-5.07(1.54)$, $-14.54(2.00)$, $-15.28(2.05)$; N: $-4.12(1.43)$, $-7.28(1.86)$, $-7.13(1.84)$. These are compiled in Table S4 (SI). Note: This is the only plot type where the color-coding does not follow the hierarchical clustering.

during this reaction. Hence, these results nicely explain the consistent presence of a catalytic effect here, within the given model approximations. Note that the best rank indicated by the arrows in the upper histogram shows an amplified case with respect to the average values, as its ESP values can be found at the edges of the main distributions of this cluster.

This corresponds to the tendency to find “outlier” solutions (needle-like lowest minima) within the clusters, observed also in

the non-neutral case of the last Section (SI). Thus, Cl is embedded in the best GOCAT more positively up to $\varphi_{\text{ESP,Cl}} = 25 \text{ kcal mol}^{-1} \text{ e}^{-1}$, N is embedded a little bit less negatively at about $\varphi_{\text{ESP,N}} = -3 \text{ kcal mol}^{-1} \text{ e}^{-1}$, and the shift at the C atom from educt to product is even increased from $\varphi_{\text{ESP,C}} = -3$ to $-18 \text{ kcal mol}^{-1} \text{ e}^{-1}$. Also all the H atoms (not shown in this Figure 7) are embedded highly symmetrically (which is not constrained in general during the GA optimization).

The broad distributions of ESP values in Figure 7 can be interpreted as a robust solution domain: As long as the ESP is within those ranges, a significant catalytic effect is to be expected, with a final barrier of about $\Delta E^{\ddagger} = 17.0 \text{ kcal mol}^{-1}$, compared with the corresponding final reaction profiles given in Figure 8

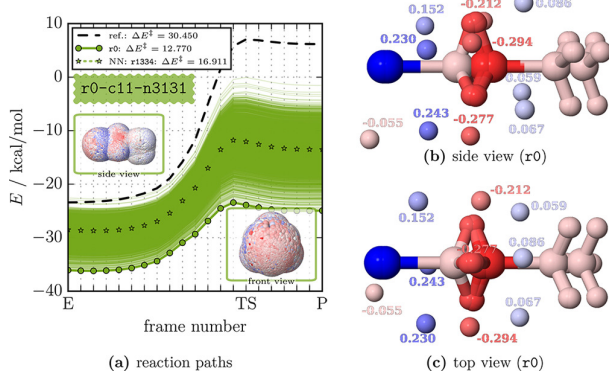


Figure 8. $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral): In a, reaction paths of the biggest cluster (c11) are shown. For illustration details, see Figure 5. Note that the overlain GOCATs in the insets of the plots are misleading, as almost a continuous embedding is shown as superposition of, e.g., each 3131 · 10 charge in the c11 case. In b, c, two different views of r0 (of c11) with values given for the partial charges. Both partial charges and the atoms of the selected core frames (E, TS, P) are colored red/blue in the ranges $[-0.537, +0.537] \text{ e}$ for charges and $[-23.578, +23.578] \text{ kcal mol}^{-1} \text{ e}^{-1}$ for ESP values. More views are given in the SI (Figure S33).

and others in the SI in Figure S31 for all the GOCATs. (Of course, these reaction paths are similar to Figure S23 for the biggest non-neutral cluster; but because of the charge neutrality constraint here, we obtain a higher resolution of this search space region). But as already mentioned above, with the increased complexity of $N_{\text{Ch}} = 10$ GOCATs, several also less symmetric or less extreme solutions are possible that are also (near) local minima within this cluster. A very specific (“outlier”) example is the best solution found, r0, that is also illustrated in Figure 8. This is one of many possible Cartesian placements of partial charges that achieve those ESP values at the main reaction atoms (Cl, N, C) but even more amplified as discussed above. Notice the high symmetry of the Cartesian placements themselves (which is not strictly necessary in very many other solutions).

As with many other, higher-rank GOCATs, also this lowest-rank one can be interpreted quite easily: The dipole moment (contact ion pair) is nicely reproduced by the GOCAT embedding of the 3 positive (near Cl atom) and the 3 negative partial charges (near C atom in the middle), respectively. Moreover, the positive charges between the attacking NH_3 group and attacked C atom facilitate the nucleophilic attack of the N atom on the slightly negative C atom, which is then fully stabilized (negatively embedded) at the P-side when the cation

has formed (thus the shift of φ_{ESP} of C from E to P in Figure 7). The remaining single charge of low value sitting at the Cl atom seems to be “mere noise”, i.e., in this $N_{\text{Ch}} = 10$ case it has to be placed somewhere, but in principle it could also vanish, i.e. set to 0.0 e during optimization automatically. Thus, this shown GOCAT is very symmetric but still not “perfect”, as at least this (tiny) 10th charge and also partially the positions and values of the other charges slightly break the symmetry.

Following this interpretation (and intuition created), one could think of even more symmetric final solutions, maybe the “real” global minimum of this search space, which would render the best rank shown above a prematurely converged one.

The “outlier” quality of the best rank (r0) discussed above partially arises because for optimal fitness a nontrivial compromise between the most important fitness ingredients has to be found: between the lowest barriers possible and the smallest gradient norm penalties. There is a clear overall trend on this fitness surface, see Figure 6 and/or Figure S30 (SI), but of the main cluster shown (c11) just a tiny fraction is found in the best (“upper left”) lowest fitness region. Looking especially at the best GOCATs there, all these show again a certain symmetry, where the mirror plane is the most important element. With that symmetry, it is also possible to reach absolute higher electrostatic potential values at the core atoms, i.e., the range between negative Cl atom embedding and positive N and especially C atom embeddings increases. With partial asymmetry, gradients (of e.g. differently embedded H atoms) increase and thus penalize most of the asymmetric GOCATs with large charge/ESP values. Since our GA has no built-in preference for symmetry, asymmetric GOCATs are much more likely than symmetric ones. Therefore, most individuals in the c11 cluster do not make it into the upper left, lowest-fitness corner. However, as discussed above, also with less (or no apparent) symmetries of the GOCAT, the main catalytic effect is already established, though to a smaller extent. One such example—the nearest neighbor to the computed cluster mean (and thus very near the centers of the distributions in Figure 7)—is given in the SI (see Figure S34).

3.4. $N_{\text{Ch}} = 10$ Stabilizing COSMO Path. The strongest restriction of the current electrostatic GOCAT model is the static preoptimized reaction path. This forces all stationary points to retain their character, despite the changing electrostatic embedding. We further discuss the influence of this assumption and how to go beyond it in the discussion and outlook section below; but even with retaining this restriction here, we still have the freedom to select any other predefined reaction path as input. We illustrate the possible benefits of other path choices for the same reaction here, by selecting a NEB path with H_2O as solvent, treated by the implicit solvent model COSMO.¹¹¹ As it is well-known for this reaction in a protic solvent, on such a path the TS is shifted toward the E frame, the barrier is lowered, and the reaction is overall exothermic—in concordance with Hammond’s postulate,¹¹² i.e., a late TS in the gas phase is shifted to the educt with increased stabilization due to the surrounding. With a slightly changed objective function, the target now is to stabilize this COSMO path—not with the COSMO implicit solvent but by our electrostatic GOCAT. This allows us to compare not only GOCAT-embedded reaction profiles to the gas-phase profile for the same path (as done above) but also the GOCAT embedding to the COSMO embedding. Moreover, this will also lead to the discussion of several other model assumptions.

So, the objective function now did not include the maximal barrier decrease, item 1., and not the TS stabilization, item 2., of Section 2 but instead tried to reach the COSMO energies themselves for each frame, and the gradient norms on the stationary points (E, TS, P) were minimized as well. (Note, however, that further tests indicated that stepping back from this setting to the original one in Section 2 leads to qualitatively very similar results as reported here.) The other ingredients were not changed (explicit weights of the terms are given in the SI again). Additionally, as in Section 3.3, the final summed charge of the GOCAT is constrained to zero.

With this COSMO path and these fitness ingredients, yet another series of global and local optimizations was performed, again collecting the results into a common database and analyzing the latter with the same tools already used above. For the biggest cluster in that database, r0-c16-n2210, the final reaction profiles are illustrated in Figure 9 (see the SI for the

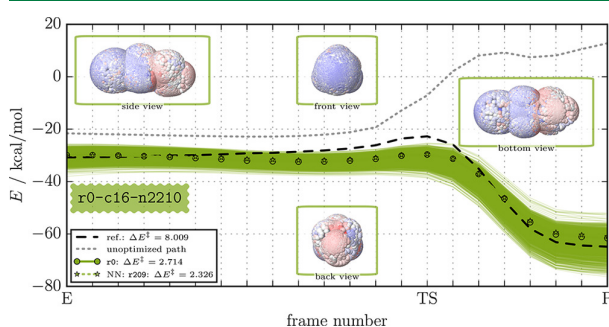


Figure 9. $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, COSMO): Reaction paths of the biggest GOCAT cluster (c16) starting with the COSMO structures but stabilized solely by partial charges (gas phase). For illustration details, see Figure 5. The “unoptimized path” is the energies of the COSMO NEB structures calculated in the gas phase without any GOCAT.

clustering plots). In general, the optimization target to reach a stabilized COSMO path (with respect to the gradients of the stationary points as well as the energies) was successful: The overall energy trend of the COSMO reaction could be reached with just 10 explicit partial charges on a vdW surface. The GOCATs are able to stabilize the differing geometric features of the structures of the COSMO case and are able to reproduce the strong stabilization of the ionic products to reach $E_{\text{P}} < E_{\text{E}}$. At the same time (and as “side effect”, as this was not an explicit optimization target in this case) the reaction barrier decreases significantly. Still, there also are a few small deviations between the COSMO and the GOCAT energy profiles; they will be discussed and explained below.

A histogram of the ESP values of this cluster is given in Figure 10. Compared to the previous case (Figure 7 in Section 3.3), the most striking differences and commonalities can be summarized as follows:

1. The overall ESP range across all GOCAT clusters is increased strongly, from $\sim[-24,+24]$ to $\sim[-84,+84]$ kcal mol $^{-1}$ e $^{-1}$.
2. N is now more negatively embedded than C (in the previous Section, C had the most negative embedding).
3. C still shows the biggest ESP shift but now between the TS and P frames. Before, the ESP shift happened between E and TS.

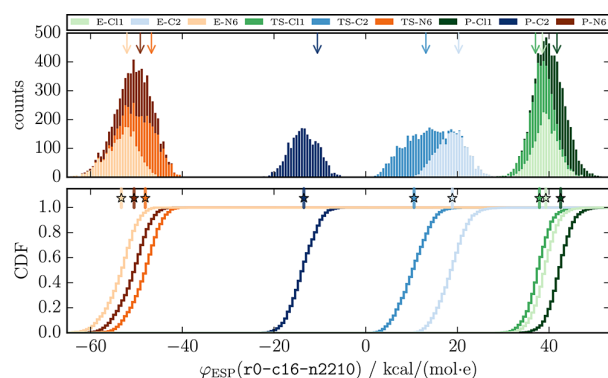


Figure 10. $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, COSMO): Stacked histogram and cumulative distribution function of φ_{ESP} for the cluster surrounding the best GOCAT. For illustration details see Figure 7. The best rank, r0, is given in Figure 11 corresponding to the arrows in the top plot. All average φ_{ESP} values (vertical bars, bottom plot) in kcal mol $^{-1}$ e $^{-1}$ for E, TS, and P frames (standard deviation in parentheses): Cl: 39.61(2.69), 37.90(2.65), 42.54(2.67); C: 18.90(3.30), 10.55(3.52), -13.49(3.06); N: -53.31(3.40), -48.09(3.27), -50.53(3.38). These are compiled in Table S4 (SI).

So still, the Cl atom at all frames (E, TS, P) is very positively embedded with small changes between the frames but now at about $\varphi_{\text{ESP,Cl}} = 40$ kcal mol $^{-1}$ e $^{-1}$. The shift of the ESP at the C atom starts now at a positive potential at E and shifts to a negative one at P between $\varphi_{\text{ESP,C}} = 19$ and -13 kcal mol $^{-1}$ e $^{-1}$, while the TS frame is also positively embedded at about $\varphi_{\text{ESP,C}} = 11$ kcal mol $^{-1}$ e $^{-1}$. The ESP at N still is negative, but this time also amplified to a region at about $\varphi_{\text{ESP,N}} = -50$ kcal mol $^{-1}$ e $^{-1}$.

The amplification part (stronger Coulomb potentials), item 1., can be traced back to the strong stabilizations of the ionic products that are necessary to reach the exothermicity of this reaction path observed in such very polar embeddings. This can also be seen in Figure 9 (compare the GOCAT profiles to the unoptimized reference energy in that plot): The possible impact of the electrostatic surrounding will always be bigger at the P frame than at the E frame because of the needed charge separation in the contact ion pair. In a crude classical picture: Along with a strong electrostatic stabilization come bigger (additional) gradients that have to be “compensated” by the internal molecular gradients, in order to become the stationary point on the effective PES. In the gas phase, this shift of stationary points had to be suppressed right away (because of the precalculated reaction path). Now the shifts of stationary points of the COSMO path are strictly necessary to be stabilized, and the surrounding can or must turn on its influence.

Furthermore, in order to discuss further subtle details, we need to consider the geometric differences between the gas phase and COSMO NEB paths (compare Figure 2 and Table 1): In the COSMO case, before NH $_3$ begins its attack onto the middle C atom, it must rotate “inwards”, since within a polar embedding the free electron pair at N is stabilized when pointing outward. Additionally, and more importantly, the stationary points on the PES are “shifted” with regard to the gas phase: The TS is “earlier” on the reaction coordinate. At the TS, the bond angles are much closer to 90°, i.e., the Walden inversion is not almost finished, as it is in the gas phase TS. Finally, in the COSMO case, the ionic products Cl $^-$ and H $_3$ NCH $_3^+$ are more widely separated, by about 0.6 Å, corresponding to a longer summed up reaction coordinate and distance between the TS and P frames. Note that at first sight these structural differences

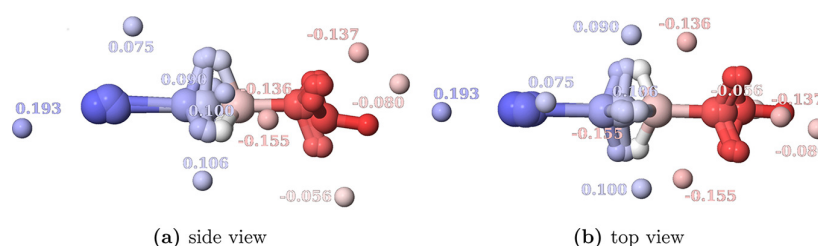


Figure 11. Single $N_{\text{Ch}} = 10$ GOCAT individual (PM7, neutral, COSMO): r_0 (of c_{16}) case with values given for the partial charges. Both partial charges and the atoms of the core frames (E, TS, P shown) are colored red/blue in the ranges $[-0.751, +0.751]$ e for charges and $[-84.142, +84.142]$ kcal mol $^{-1}$ e $^{-1}$ for ESP values. More views are given in the SI (Figure S44).

may appear to be small, but they are accompanied by significant gradient changes: By construction, the TS of the COSMO path has a gradient value of zero in the COSMO reference calculation, and it has a small gradient in reasonable GOCATs. However, in the gas phase it has a gradient of 78.5 kcal mol $^{-1}$ Å $^{-1}$, well beyond the allowed tolerance. Hence, the (necessary) gradient norm penalties prevented the GOCAT optimizations in Section 3.3 to achieve reaction profiles of the COSMO-like kind shown in the present subsection.

In Section 3.3, within the GOCAT case starting with the gas phase NEB path, we saw a boundary of positive and negative charges between Cl and C (see Figure 8 and the text above). The second boundary to the positive partial charges between C and N leads to the result that the biggest (negative) ESP value was found at C (at the TS and P frames). Now within COSMO of this Section, there is just a single boundary of partial charges: Positive charges are sitting either at Cl or besides C (at the E frame and also at the TS frame because of the TS shift), and negative ones are sitting at the N side, see Figure 11, where the very symmetric best rank of the single cluster discussed in this Section is illustrated. Thus, this observation geometrically reflects item 2. and item 3.: The ESP values are positive at the Cl site and monotonically become negative at the N site. The shift of ESP at the C atom arises between TS and P and not between E and TS, since the TS frame has moved toward the E frame.

A further illustration of our electrostatic GOCAT embeddings can be obtained by comparison to the COSMO embedding itself. For this purpose, this COSMO embedding is translated into a partial charge surrounding, as illustrated in Figure 12. Per construction in COSMO as apparent surface charge model, partial charges are shown as representatives of the quadrature over the discrete surface tesserae of constant charge density on the cavity's surface. [Due to some numerical problems and/or a principally different surface construction mechanism in MOPAC compared to our framework, this surface was constructed on a solvent accessible surface (SAS) with a slightly decreased vdW radius of the probe molecule: This is just for convenience of illustration. The final ESP values as well as COSMO properties were not influenced by this modification.] So, in a perfect conductor the apparent electrostatic potentials at the cavity's surface are exactly compensated by a potential due to the induced (counter-)charges on the cavity's surface that shield (screen) the solute (and afterward rescaled with a simple screening factor to reach the finite dielectric constant of water).^{111,113}

This pure COSMO embedding can now serve as a reference for the interpretation of our GOCATs. We could think of the cluster of GOCATs treated in this Section and also of the single GOCAT shown in Figure 11 as globally best and coarsened approximation to COSMO. (Of course, this is for illustrative

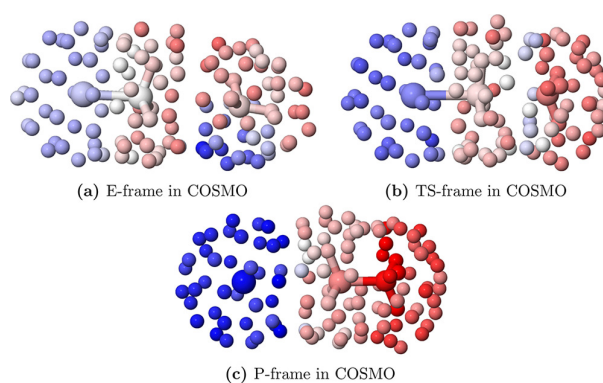


Figure 12. Side views of E, TS, and P stationary point structures on PM7 with COSMO (H $_2$ O). The quadrature used within COSMO is illustrated by explicit partial charges. The latter are colored red/blue in the range $[-0.03, +0.03]$ e, and the φ_{ESP} values are also calculated and colored in the range $[-92.92, +92.92]$ kcal mol $^{-1}$ e $^{-1}$. Explicit ESP values in kcal mol $^{-1}$ e $^{-1}$ for E, TS, and P frames: Cl: 13.95, 38.39, 85.37; C: -2.83, -4.17, -33.45; N: -7.38, -37.21, -82.85. These are compiled in Table S4 (SI).

purposes only, and this comparison ignores several technical details, e.g., using a vdW surface instead of a solvent accessible surface or using a minimal distance threshold between the charges instead of near uniformly distributed charges on segments or tesserae.)

Comparing the GOCAT solutions (Figure 11 and ESP values in Figure 10) with the pure COSMO case in Figure 12, we see again ESP values with the following trend: $\varphi_{\text{ESP,Cl}} > \varphi_{\text{ESP,C}} > \varphi_{\text{ESP,N}}$ with positive embeddings for Cl and negative ones for every other atom, including C and N (but also including each separate H atom, compare with Table S4 in the SI), which was item 2. above. As easily anticipated, during the reaction starting with neutral educts and ending with charged products, the absolute ESP values just increase (the total range of ESP values enlarges), to better screen (and solvate) the products.

The neutral educt frame in COSMO has very small ESP values (and correspondingly small charges), while also some positive (blue) charges can be found close to the free electron pair of NH $_3$. Besides the fact of this stabilization (which is also the reason for lifting the C_{3v} symmetry for the educts), this obviously explicitly addresses individual atoms better and is of course due to the SAS construction principle (i.e., almost uniform segment-based embedding which also leads to many but very small partial charges). Still, the accumulated electrostatic potential at N is negative throughout. On the product side, the charges turn up their impact and their electrostatic potential, to fully stabilize the charged products. In contrast, in the GOCAT model we have

just a few charges (with higher absolute values) that try to find a compromise: As the COSMO NEB is fixed and not relaxed for each separate GOCAT in our GA, the common cavity (vdW surface) is the same for each frame and not different for each frame like in the COSMO case. Moreover, a partial charge that tries to stabilize the product frame a great deal (which is the most important handle, of course) will influence the educts, too. In other words, there is just *one* GOCAT surrounding the complete reaction and not a separate one for each NEB frame. So, this does not allow a “dynamic” change or reaction of the surrounding with regard to the changed electrostatic situation of the reaction frames. In contrast, in COSMO it is assumed that the surrounding is instantaneously following the reaction in a frame-wise manner (however small the steps from one frame to the other may be) and that the bidirectional influence between the surroundings and the geometric location of the path has been iterated to self-consistency. The positive charges in the NH₃ educt frame missing in the GOCAT case as well as significantly higher ESP values at the educt is thus one reason for some subtle intermediate minima between E and TS (in some GOCATs, the energy is not strictly monotonically increasing from E to TS, see Figure 9). Additionally, the overall “compromise” leads to too much stabilization at the TS frame and to less stabilization at the P frame due to finally smaller absolute φ_{ESP} values (compare the best rank r_0 with the COSMO reference in Figure 9): $\Delta E_{\text{E-GOCAT,stab.}} > \Delta E_{\text{E-COSMO,stab.}} \wedge \Delta E_{\text{TS-GOCAT,stab.}} < \Delta E_{\text{TS-COSMO,stab.}} \wedge \Delta E_{\text{P-GOCAT,stab.}} > \Delta E_{\text{P-COSMO,stab.}}$. Finally, the one and only shift of ESP is observed at the C atom, as this exhibits the biggest Cartesian move from E to P (within the maximally compact superposed reaction frames) and can thus be influenced differently, item 3. above.

In summary, keeping the technical differences discussed above in mind, our electrostatic GOCATs can be reinterpreted as coarsened, nondynamic approximations to COSMO embeddings, for this specific Menshutkin reaction showing this clear charge separation. As shown above, sufficient similarities exist between our best GOCATs and those COSMO embeddings to take the latter as further support that the former work correctly and successfully.

4. FURTHER DISCUSSION OF THE MODEL

The present contribution is a proof-of-concept study, containing several fundamental model approximations, and even after future elimination of these approximations several further steps need to be taken to arrive at a truly comprehensive inverse design of catalysts. To avoid miscategorizations of our work, these issues are addressed in Section 4.2 below. Before that, in Section 4.1 we briefly digress toward possible connections between different levels of theory.

4.1. Level of Theory. In Section 3, we have deliberately limited ourselves to the PM7 level of theory. This choice was motivated by the possibility to do many exploratory calculations quickly, while still retaining a quantum-chemical foundation. Of course, however, our GOCAT concept is completely independent of any particular level of theory. For example, we could repeat the present work, using e.g. MP2 or DFT with a suitable functional. This would increase computation times roughly by a factor of 100. However, after having established with the present work how to set up the necessary algorithmic ingredients (e.g., how to choose the fitness function) and by exploiting the built-in trivial parallelism, such a project is now realistic.

To explore possible difficulties and to generate a preliminary glimpse at possible results of such a larger-scale study, we have already done a stripped-down version of it, using PBE0/def2-TZVP in the ORCA program suite.

Note that there are subtle differences in details: For PM7 as implemented in MOPAC, just the φ_{ESP} at the atomic positions is needed;⁸⁵ usually, one would need also a semiempirical parametrization for the electrostatic potential on the same footing.^{114,115} In contrast, for DFT in ORCA, the full scalar-field Coulomb potential is used during the SCF, as induced by additional “external” point charges (i.e., analogous to new fractional “nuclear core” positions with nuclear attraction/repulsion but no additional bases). Thus, in these DFT calculations, each separate charge *position* is important, not just the final summed Coulomb potential at the core atoms.

Because of the increase of computational time by about 2 orders of magnitude, we resorted to several time-savers: (1) We used the PM7 GOCATs as initial seeds, instead of random seeds; (2) we only did two separate GA runs, with an order of magnitude less GA iterations; and (3) we locally optimized only part of the GOCATs in the resulting, joined database (not all of them). From these clustered results, we briefly examine only the biggest cluster here.

In Figure 13 both reaction profiles are shown as well as the best DFT optimized so far. Again, also on this higher

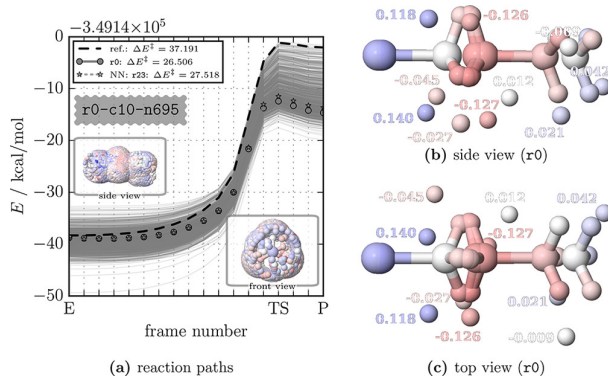


Figure 13. $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral): In a: Reaction paths of the biggest cluster (c10) are shown; other illustration details are similar to Figure 5. In b, c: Two views on the best GOCAT individual r_0 in that cluster. Both partial charges and the atoms of the core frames (E, TS, P) are colored red/blue in the ranges $[-0.645, +0.645]$ e for charges and $[-43.675, +43.675]$ kcal mol⁻¹ e⁻¹ for ESP values, explicitly: Cl: 9.61, 9.94, 9.75; C: 0.97, -10.80, -12.09; N: -0.78, -6.34, -6.47.

level of theory, we are able to get significant catalytic effects of almost $\Delta\Delta E^\ddagger = 10$ kcal mol⁻¹ with respect to the (DFT) gas-phase profile. Because of the time-savers listed above, some more spread is still visible; but the final best GOCAT found so far is qualitatively similar to the PM7 result (compare the GOCAT charges and φ_{ESP} values in Figure 13 with Figure 8, compiled also in Table S4 (SI)). Still, the charge values are a bit smaller corresponding also to a smaller ESP at the core atoms.

Besides that, already geometric translations of PM7 GOCATs to PBE0 without any optimization at all led to a high percentage of GOCATs that also show at least a minor catalytic effect as well as a significantly better starting fitness of the complete database compared to randomly initiated GOCATs.

However, most final separate best GOCATs we found were not as easily interpretable as the ones on PM7, i.e., the strict separation of positive and negative charges in the Cartesian domain and the high symmetry was not (yet) found. We ascribe this to the (prematurely converged) smaller DFT database. Some more impressions on this level of theory are given in the SI, including also an example of relaxed stationary points within a GOCAT (a posteriori lifting the assumption of a fixed reaction path).

4.2. Model Approximations. In Table 2 both some principal model assumptions and important steps are illustrated

Table 2. Overview of Important Ingredients To Reach a Better Representation of a Catalyst (Not Exhaustive)

real catalytic embedding		electrostatic GOCAT model
“relaxed” reaction path	⇔	fixed/preoptimized reaction path
multistep reaction path	⇔	single step (one TS)
catalyst may enable new mechanism	⇔	given, fixed mechanism
effects of catalyst–substrate dis-/association	⇔	catalyst–substrate complex only
arbitrary substrate–catalyst interactions	⇔	QM/MM electrostatic embedding
concrete real catalyst	⇔	abstract embedding
ZPE, thermal energy, free enthalpy (including entropy)	⇔	potential energy
possible catalyst reorganization	⇔	one GOCAT for the whole reaction path

as to how the proposed simple electrostatic GOCAT model could be extended to approach a more comprehensive catalysis design process.

As illustrated in Section 3 by the marked difference between gas-phase-based and COSMO-based GOCATs, a significant limitation of our GOCAT model in its present, early development stage is keeping the reaction path fixed throughout, both as a path in coordinate space and as a reaction profile in energy with (approximately) retained localizations of stationary points. In the real world, it is likely that the catalytic embedding changes the reaction path in both of these respects (i.e., imagine changed mechanisms at a heterogeneous scaffold, within an enzyme or at a transition-metal complex; in the context of the present reaction, also an S_N2 to S_N1 shift seems likely). Thus, a full reaction-path relaxation within the GOCAT embedding during GOCAT optimization would be advantageous: This would eliminate the need for additional gradient thresholds and would allow for bigger GOCAT impacts. In the examples above, we would have reached COSMO-like GOCATs despite starting out from the gas-phase reaction path. However, a price to pay for this advantage is a substantial increase in computational effort. Work along these lines is in progress.

As already hinted at in the Introduction, many people understand “catalysis” (at transition-metal centers, in enzymes, etc.) as complicated, multistep reaction sequences, which can be formulated as multistep reaction cycles in which the catalyst is finally regenerated in its original form. Here we have focused on lowering the activation energy of a single-step reaction. This can either be understood as another (also frequently used) interpretation of “catalysis”, or we can of course argue that concatenation of several (catalyzed) single steps eventually leads to multistep sequences. Usually, even in longer multistep sequences a first approach is to narrow down which TS is the actual bottleneck. Then the inverse design problem can be

reduced to this step, which again is a single one. For more automation and less user intervention, frameworks for dynamic reaction networks and automatic optimization of reaction paths exist (for some examples, see refs 72–77 and references therein), which could be combined with our GOCAT approach.

However, the influence of a catalyst on a reaction may be even more extreme: It could make new reaction mechanisms possible that are inconceivable in the absence of this particular catalyst and hence cannot be part of any preconceived single- or multistep reaction pathway. Clearly, our present GOCAT setup is very far away from such phenomena. In principle, however, a suitable combination of multistep GOCAT design and automatic reaction network exploration is conceivable for such advanced purposes.

Yet another limitation of our present setup is its exclusive focus on the situation in which catalyst and substrate are in their tightest association. Arguably, this is the most important, central situation of catalysis, but equally clearly it is not the whole story. How the reactants and the catalyst get into this catalytic pose, from an initial infinite separation, and how they dissociate again is also important and has influence on the whole catalytic process, as witnessed e.g. by studies on how proteins “channel” their substrates to, between, and away from their active sites.¹¹⁶

As also hinted at in the Introduction, there is an ongoing controversy on whether catalysis involves various substrate-catalyst interactions (including H-bonds, vdW interactions, etc.) or only electrostatic ones. Our GOCAT model currently represents only the latter but can be extended to also include the former. Indeed, besides partial charges, we aim at introducing other (abstract) interaction groups for also treating vdW interactions, H-bond donors/acceptors, etc.

Of course, a defining feature of our GOCATs is their strong abstraction. As already seen in these very first examples, these abstract catalytic fields can help to analyze and understand known catalytic effects. Similarly, it could help chemists to manually design new, as yet unknown catalysts for a given reaction. However, actually translating a GOCAT into a real catalyst clearly requires substantial further effort. In general, the translation back to a real molecular embedding will likely always introduce some “translation error”, depending on how well chemically realizable the embedding is. An alternative route toward a real chemical catalyst would be a global optimization with real (capped) interaction entities directly from the beginning, i.e., global optimization of a “theozyme” with a more complex objective function than just the TS energy.

Furthermore, real catalysts operate at finite temperature on free enthalpies, and clearly the association between substrate and catalyst has an effect on entropy. In contrast, our GOCATs currently operate only on bare potential energy values. While adding in some statistical thermodynamics corrections within the usual harmonic approximations is easily possible, taking into account all necessary entropy and anharmonicity effects is challenging.

As noted in Section 3.4, a substantial difference between the COSMO embedding and our GOCAT was that the latter provides one compromise embedding across all points on the reaction path, while the former by hypothesis instantly adapts to all changes in the reaction center along the reaction path. It would be possible to extend the GOCAT concept toward multiple GOCATs along the reaction path, with restraints to limit changes between successive ones to what is physically possible. This would then model a GOCAT with conformational

freedom and changing polarization states (charges), contributing to the reaction coordinate.

5. CONCLUSIONS

Finding an efficient catalyst for a given chemical reaction can be understood as a molecular design task. As a first stage on a longer journey toward this goal, we have presented our “Globally Optimal Catalyst” (GOCAT) framework. It introduces an abstract representation of the catalytic surrounding that is described by a strictly electrostatic model in the present contribution, to reduce and simplify the search space. Within this abstract space, we then search for optimal point charge arrangements, using a nondeterministic global search (Genetic Algorithms) and a multicomponent objective function in which “barrier decrease” is only one of several ingredients.

As an illustrative example, we have applied this first-stage framework to the prototype Menshutkin S_N2 reaction, using the semiempirical PM7 level of theory to enable extensive algorithm testing and varying numbers of point charges (from 1 to 10). Both for the gas-phase and for the solution-phase reaction path, we could find point-charge surroundings with significant catalytic effects. Additionally, our GOCATs for the solution-phase path (but without a solvent model present) clearly mimic the surrounding that the continuum solvent model provides.

To make the voluminous databases of GOCAT candidates that our GA runs produce accessible to comparative inspection and interpretation, we employed dimension reduction techniques as well as unsupervised Machine Learning (Clustering). For the presented example reaction, our findings indicate that already in its first development stage this GOCAT approach has the potential both to illuminate the decisive interactions behind electrostatic catalytic effects and to aid in finding new catalytic embeddings.

We have concluded this Article with a discussion of further development steps that need to be taken to arrive at the ultimate goal of fully automatic design of optimized real-world catalysts for any given reaction.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00151.

Complementary figures and some statistics in tables for each case in the article; figure set for $N_{Ch} = 3$ missing in this article; more information on the used two objective functions and a GOCAT optimization without a strict TS gradient norm check within the objective function (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hartke@pctc.uni-kiel.de.

ORCID

Mark Dittner: 0000-0001-5513-7053

Bernd Hartke: 0000-0001-8480-0862

Funding

M.D. gratefully acknowledges financial support by a fellowship of the German Fonds of the Chemical Industry (FCI). B.H. thanks the German Science Foundation DFG for financial support via grant Ha2498/16-1.

Notes

The authors declare no competing financial interest.

■ REFERENCES

- (1) Weymuth, T.; Reiher, M. Inverse Quantum Chemistry: Concepts and Strategies for Rational Compound Design. *Int. J. Quantum Chem.* **2014**, *114*, 823–837.
- (2) von Lilienfeld, O. A. Towards the Computational Design of Compounds from First Principles. In *Many-Electron Approaches in Physics, Chemistry and Mathematics: A Multidisciplinary View*; Bach, V., Delle Site, L., Eds.; Mathematical Physics Studies; Springer International Publishing: Cham, 2014; pp 169–189, DOI: 10.1007/978-3-319-06379-9_9.
- (3) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432*, 823–823.
- (4) von Lilienfeld, O. A. First Principles View on Chemical Compound Space: Gaining Rigorous Atomistic Control of Molecular Properties. *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689.
- (5) Besenbacher, F.; Chorkendorff, I.; Clausen, B. S.; Hammer, B.; Molenbroek, A. M.; Nørskov, J. K.; Stensgaard, I. Design of a Surface Alloy Catalyst for Steam Reforming. *Science* **1998**, *279*, 1913–1915.
- (6) Marder, S. R.; Beratan, D. N.; Cheng, L.-T. Approaches for Optimizing the First Electronic Hyperpolarizability of Conjugated Organic Molecules. *Science* **1991**, *252*, 103–106.
- (7) Kuhn, C.; Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **1996**, *100*, 10595–10599.
- (8) Ceder, G.; Chiang, Y.-M.; Sadoway, D. R.; Aydinol, M. K.; Jang, Y.-L.; Huang, B. Identification of Cathode Materials for Lithium Batteries Guided by First-Principles Calculations. *Nature* **1998**, *392*, 694–696.
- (9) Krausbeck, F.; Sobez, J.-G.; Reiher, M. Stabilization of Activated Fragments by Shell-Wise Construction of an Embedding Environment. *J. Comput. Chem.* **2017**, *38*, 1023–1038.
- (10) De Vleeschouwer, F.; Geerlings, P.; Proft, F. D. Molecular Property Optimizations with Boundary Conditions through the Best First Search Scheme. *ChemPhysChem* **2016**, *17*, 1414–1424.
- (11) Franceschetti, A.; Zunger, A. The Inverse Band-Structure Problem of Finding an Atomic Configuration with given Electronic Properties. *Nature* **1999**, *402*, 60–63.
- (12) Carstensen, O. N.; Dieterich, J. M.; Hartke, B. Design of Optimally Switchable Molecules by Genetic Algorithms. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2903–2910.
- (13) von Lilienfeld, O. A.; Lins, R. D.; Rothlisberger, U. Variational Particle Number Approach for Rational Compound Design. *Phys. Rev. Lett.* **2005**, *95*, 153002.
- (14) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing Molecules by Optimizing Potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228–3232.
- (15) Mueller, T.; Kusne, A. G.; Ramprasad, R. Machine Learning in Materials Science: Recent Progress and Emerging Applications. In *Reviews in Computational Chemistry*, 1st ed.; Parrill, A. L., Lipkowitz, K. B., Eds.; John Wiley & Sons, Inc.: 2016; Vol. 29, pp 186–273, DOI: 10.1002/9781119148739.ch4.
- (16) Jadrich, R. B.; Lindquist, B. A.; Truskett, T. M. Recent Advances in Accelerated Discovery through Machine Learning and Statistical Inference. 2017, arXiv:1706.05405. arXiv.org ePrint archive. <https://arxiv.org/abs/1706.05405v1> (accessed Apr 15, 2018).
- (17) Huan, T. D.; Mannodi-Kanakithodi, A.; Ramprasad, R. Accelerated Materials Property Predictions and Design Using Motif-Based Fingerprints. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 014106.
- (18) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268.
- (19) Rangarajan, S.; Maravelias, C. T.; Mavrikakis, M. Sequential-Optimization-Based Framework for Robust Modeling and Design of Heterogeneous Catalytic Systems. *J. Phys. Chem. C* **2017**, *121*, 25847–25863.
- (20) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337–1344.
- (21) Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L.; Aspuru-Guzik, A. Optimizing Distributions over Molecular Space. An

Objective-Reinforced Generative Adversarial Network for Inverse-Design Chemistry (ORGANIC). 2017, chemRxiv:5309668. chemRxiv.org ePrint archive. https://chemrxiv.org/articles/ORGANIC_1_pdf/5309668 (accessed Apr 15, 2018).

(22) Poree, C.; Schoenebeck, F. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Acc. Chem. Res.* **2017**, *50*, 605–608.

(23) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chem., Int. Ed.* **2013**, *52*, 5700–5725.

(24) Occhipinti, G.; Koudriavtsev, V.; Törnroos, K. W.; Jensen, V. R. Theory-Assisted Development of a Robust and Z-Selective Olefin Metathesis Catalyst. *Dalton Trans.* **2014**, *43*, 11106–11117.

(25) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37.

(26) Na, J.; Houk, K. Predicting Antibody Catalyst Selectivity from Optimum Binding of Catalytic Groups to a Hapten. *J. Am. Chem. Soc.* **1996**, *118*, 9204–9205.

(27) Tantillo, D. J.; Jiangang, C.; Houk, K. N. Theozymes and Compuzymes: Theoretical Models for Biological Catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2*, 743–750.

(28) Dahiyat, B. I.; Mayo, S. L. Protein Design Automation. *Protein Sci.* **1996**, *5*, 895–903.

(29) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D. New Algorithms and an in Silico Benchmark for Computational Enzyme Design. *Protein Sci.* **2006**, *15*, 2785–2794.

(30) Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S. Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution. *J. Mol. Biol.* **2011**, *407*, 391–412.

(31) Sokalski, W. A. Theoretical model for exploration of catalytic activity of enzymes and design of new catalysts: CO₂ hydration reaction. *Int. J. Quantum Chem.* **1981**, *20*, 231–240.

(32) Sokalski, W. A. Nonempirical Modeling of the Static and Dynamic Properties of the Optimum Environment for Chemical Reactions. *J. Mol. Struct.: THEOCHEM* **1986**, *138*, 77–87.

(33) Sokalski, W. A. The Physical Nature of Catalytic Activity Due to the Molecular Environment in Terms of Intermolecular Interaction Theory: Derivation of Simplified Models. *J. Mol. Catal.* **1985**, *30*, 395–410.

(34) Szeferczyk, B.; Mulholland, A. J.; Ranaghan, K. E.; Sokalski, W. A. Differential Transition-State Stabilization in Enzyme Catalysis: Quantum Chemical Analysis of Interactions in the Chorismate Mutase Reaction and Prediction of the Optimal Catalytic Field. *J. Am. Chem. Soc.* **2004**, *126*, 16148–16159.

(35) Gérczei, T.; Asbóth, B.; Náráy-Szabó, G. Conservative Electrostatic Potential Patterns at Enzyme Active Sites: The Anion-Cation-Anion Triad. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 310–315.

(36) Barbany, M.; Gutiérrez-de Terán, H.; Sanz, F.; Villà-Freixa, J.; Warschel, A. On the Generation of Catalytic Antibodies by Transition State Analogues. *ChemBioChem* **2003**, *4*, 277–285.

(37) Kangas, E.; Tidor, B. Electrostatic Complementarity at Ligand Binding Sites: Application to Chorismate Mutases. *J. Phys. Chem. B* **2001**, *105*, 880–888.

(38) Beker, W.; van der Kamp, M. W.; Mulholland, A. J.; Sokalski, W. A. Rapid Estimation of Catalytic Efficiency by Cumulative Atomic Multipole Moments: Application to Ketosteroid Isomerase Mutants. *J. Chem. Theory Comput.* **2017**, *13*, 945–955.

(39) Dyguda-Kazimierowicz, E.; Sokalski, W.; Leszczyński, J. Non-Empirical Study of the Phosphorylation Reaction Catalyzed by 4-Methyl-5- β -Hydroxyethylthiazole Kinase: Relevance of the Theory of Intermolecular Interactions. *J. Mol. Model.* **2007**, *13*, 839–849.

(40) Szarek, P.; Dyguda-Kazimierowicz, E.; Tachibana, A.; Sokalski, W. A. Physical Nature of Intermolecular Interactions within cAMP-Dependent Protein Kinase Active Site: Differential Transition State

Stabilization in Phosphoryl Transfer Reaction. *J. Phys. Chem. B* **2008**, *112*, 11819–11826.

(41) Chudyk, E. I.; Dyguda-Kazimierowicz, E.; Langner, K. M.; Sokalski, W. A.; Lodola, A.; Mor, M.; Sirirak, J.; Mulholland, A. J. Nonempirical Energetic Analysis of Reactivity and Covalent Inhibition of Fatty Acid Amide Hydrolase. *J. Phys. Chem. B* **2013**, *117*, 6656–6666.

(42) Bhowmick, A.; Sharma, S. C.; Head-Gordon, T. The Importance of the Scaffold for de Novo Enzymes: A Case Study with Kemp Eliminase. *J. Am. Chem. Soc.* **2017**, *139*, 5793–5800.

(43) Vaissier, V.; Sharma, S. C.; Schaettle, K.; Zhang, T.; Head-Gordon, T. Computational Optimization of Electric Fields for Improving Catalysis of a Designed Kemp Eliminase. *ACS Catal.* **2018**, *8*, 219–227.

(44) Weymuth, T.; Reiher, M. Gradient-Driven Molecule Construction: An Inverse Approach Applied to the Design of Small-Molecule Fixating Catalysts. *Int. J. Quantum Chem.* **2014**, *114*, 838–850.

(45) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: New York, 1989.

(46) Weise, T. *Global Optimization Algorithms - Theory and Application*, 3rd ed.; 2011. <http://www.it-weise.de/projects/bookNew.pdf> (accessed Apr 15, 2018).

(47) Weise, T.; Chiong, R.; Táng, K. Evolutionary Optimization: Pitfalls and Booby Traps. *J. Comp. Sci. Technol.* **2012**, *27*, 907–936.

(48) Hartke, B. Global Cluster Geometry Optimization by a Phenotype Algorithm with Niches: Location of Elusive Minima, and Low-Order Scaling with Cluster Size. *J. Comput. Chem.* **1999**, *20*, 1752–1759.

(49) Dieterich, J. M.; Hartke, B. OGOLEM: Global Cluster Structure Optimisation for Arbitrary Mixtures of Flexible Molecules. A Multiscale, Object-Oriented Approach. *Mol. Phys.* **2010**, *108*, 279–291.

(50) Dieterich, J. M.; Hartke, B. Composition-Induced Structural Transitions in Mixed Lennard-Jones Clusters: Global Reparametrization and Optimization. *J. Comput. Chem.* **2011**, *32*, 1377–1385.

(51) Dittner, M.; Hartke, B. Conquering the Hard Cases of Lennard-Jones Clusters with Simple Recipes. *Comput. Theor. Chem.* **2017**, *1107*, 7–13.

(52) Dieterich, J. M.; Hartke, B. A Graph-Based Short-Cut to Low-Energy Structures. *J. Comput. Chem.* **2014**, *35*, 1618–1620.

(53) Dieterich, J. M.; Hartke, B. Improved Cluster Structure Optimization: Hybridizing Evolutionary Algorithms with Local Heat Pulses. *Inorganics* **2017**, *5*, 64.

(54) Dieterich, J. M.; Hartke, B. Empirical Review of Standard Benchmark Functions Using Evolutionary Global Optimization. *Appl. Math.* **2012**, *03*, 1552–1564.

(55) Dieterich, J. M.; Gerstel, U.; Schröder, J.-M.; Hartke, B. Aggregation of Kanamycin A: Dimer Formation with Physiological Cations. *J. Mol. Model.* **2011**, *17*, 3195–3207.

(56) Larsson, H. R.; van Duin, A. C. T.; Hartke, B. Global Optimization of Parameters in the Reactive Force Field ReaxFF for SiOH. *J. Comput. Chem.* **2013**, *34*, 2178–2189.

(57) Dittner, M.; Müller, J.; Aktulga, H. M.; Hartke, B. Efficient Global Optimization of Reactive Force-Field Parameters. *J. Comput. Chem.* **2015**, *36*, 1550–1561.

(58) Müller, J.; Hartke, B. reaxFF Reactive Force Field for Disulfide Mechanochemistry, Fitted to Multireference Ab Initio Data. *J. Chem. Theory Comput.* **2016**, *12*, 3913–3925.

(59) Warschel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.

(60) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(61) Politzer, P.; Murray, J. S.; Clark, T. Mathematical Modeling and Physical Reality in Noncovalent Interactions. *J. Mol. Model.* **2015**, *21*, 52.

- (62) Clark, T. Halogen Bonds and σ -Holes. *Faraday Discuss.* **2017**, *203*, 9–27.
- (63) Clark, T. Polarization, Donor–acceptor Interactions, and Covalent Contributions in Weak Interactions: A Clarification. *J. Mol. Model.* **2017**, *23*, 297.
- (64) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* **2006**, *106*, 3210–3235.
- (65) Kamerlin, S. C. L.; Warshel, A. At the Dawn of the 21st Century: Is Dynamics the Missing Link for Understanding Enzyme Catalysis? *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1339–1375.
- (66) Warshel, A.; Bora, R. P. Perspective: Defining and Quantifying the Role of Dynamics in Enzyme Catalysis. *J. Chem. Phys.* **2016**, *144*, 180901.
- (67) Zhang, X.; Houk, K. N. Why Enzymes Are Proficient Catalysts: Beyond the Pauling Paradigm. *Acc. Chem. Res.* **2005**, *38*, 379–385.
- (68) Gao, J.; Ma, S.; Major, D. T.; Nam, K.; Pu, J.; Truhlar, D. G. Mechanisms and Free Energies of Enzymatic Reactions. *Chem. Rev.* **2006**, *106*, 3188–3209.
- (69) Jiménez, A.; Clapés, P.; Crehuet, R. A Dynamic View of Enzyme Catalysis. *J. Mol. Model.* **2008**, *14*, 735–746.
- (70) Christofferson, A.; Zhao, L.; Pei, Q. Dynamic Simulations as a Complement to Experimental Studies of Enzyme Mechanisms. *Adv. Protein Chem. Struct. Biol.* **2012**, *87*, 293–335.
- (71) Závodszy, P.; Hajdú, I. Evolution of the Concept of Conformational Dynamics of Enzyme Functions over Half of a Century: A Personal View. *Biopolymers* **2013**, *99*, 263–269.
- (72) Zimmerman, P. M. Automated Discovery of Chemically Reasonable Elementary Reaction Steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.
- (73) Habershon, S. Automated Prediction of Catalytic Mechanism and Rate Law Using Graph-Based Reaction Path Sampling. *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798.
- (74) Sameera, W. M. C.; Maeda, S.; Morokuma, K. Computational Catalysis Using the Artificial Force Induced Reaction Method. *Acc. Chem. Res.* **2016**, *49*, 763–773.
- (75) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J. Chem. Theory Comput.* **2017**, *13*, 5780–5797.
- (76) Simm, G. N.; Reiher, M. Context-Driven Exploration of Complex Chemical Reaction Networks. *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119.
- (77) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1354.
- (78) Dieterich, J. M. ogolem.org homepage. <https://www.ogolem.org/> (accessed Apr 15, 2018).
- (79) Bandow, B.; Hartke, B. Larger Water Clusters with Edges and Corners on Their Way to Ice: Structural Trends Elucidated with an Improved Parallel Evolutionary Algorithm. *J. Phys. Chem. A* **2006**, *110*, 5809–5822.
- (80) Dieterich, J. M.; Hartke, B. Error-Safe, Portable, and Efficient Evolutionary Algorithms Implementation with High Scalability. *J. Chem. Theory Comput.* **2016**, *12*, 5226–5233.
- (81) Spenke, F.; Balzer, K.; Frick, S.; Hartke, B.; Dieterich, J. M. Adaptive Parallelism with RMI: Idle High-Performance Computing Resources Can Be Completely Avoided. 2018, arXiv:1801.07184. arXiv.org ePrint archive. <https://arxiv.org/abs/1801.07184> (accessed Apr 15, 2018).
- (82) Powell, M. J. *The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives*; Technical Report NA2009/06; Department of Applied Mathematics and Theoretical Physics, Cambridge, 2009.
- (83) Stewart, J. J. P. *MOPAC2016*; Stewart Computational Chemistry: Colorado Springs, CO, USA, 2016.
- (84) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1–32.
- (85) Plotnikov, N. V.; Warshel, A. Exploring, Refining, and Validating the Paradynamics QM/MM Sampling. *J. Phys. Chem. B* **2012**, *116*, 10342–10356.
- (86) Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 73–78.
- (87) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0Model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (88) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
- (89) Deaven, D. M.; Ho, K. M. Molecular Geometry Optimization with a Genetic Algorithm. *Phys. Rev. Lett.* **1995**, *75*, 288–291.
- (90) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer Series in Statistics; Springer New York: New York, NY, 2009; DOI: 10.1007/978-0-387-84858-7.
- (91) Jørgensen, M. S.; Groves, M. N.; Hammer, B. Combining Evolutionary Algorithms with Clustering toward Rational Global Structure Optimization at the Atomic Scale. *J. Chem. Theory Comput.* **2017**, *13*, 1486–1493.
- (92) Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.
- (93) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.
- (94) Bahn, S.; Jacobsen, K. An Object-Oriented Scripting Interface to a Legacy Electronic Structure Code. *Comput. Sci. Eng.* **2002**, *4*, 56–66.
- (95) Bitzek, E.; Koskinen, P.; Gähler, F.; Moseler, M.; Gumbusch, P. Structural Relaxation Made Simple. *Phys. Rev. Lett.* **2006**, *97*, 170201.
- (96) Sheppard, D.; Terrell, R.; Henkelman, G. Optimization Methods for Finding Minimum Energy Paths. *J. Chem. Phys.* **2008**, *128*, 134106.
- (97) Melander, M.; Laasonen, K.; Jónsson, H. Removing External Degrees of Freedom from Transition-State Search Methods Using Quaternions. *J. Chem. Theory Comput.* **2015**, *11*, 1055–1062.
- (98) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
- (99) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.
- (100) von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier Series of Atomic Radial Distribution Functions: A Molecular Fingerprint for Machine Learning Models of Quantum Chemical Properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084–1093.
- (101) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
- (102) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.
- (103) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- (104) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 10–20.
- (105) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (106) Jain, A. K. Data Clustering: 50 Years beyond K-Means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666.

(107) Abboud, J.-u. M.; Notario, R.; Bertrán, J.; Solà, M. One Century of Physical Organic Chemistry: The Menshutkin Reaction. In *Progress in Physical Organic Chemistry*; Taft, R. W., Ed.; John Wiley & Sons, Inc.: 1993; pp 1–182, DOI: 10.1002/9780470171981.ch1.

(108) Li, Y.; Hartke, B. Assessing Solvation Effects on Chemical Reactions with Globally Optimized Solvent Clusters. *ChemPhysChem* **2013**, *14*, 2678–2686.

(109) Giacinto, P.; Zerbetto, F.; Bottoni, A.; Calvaresi, M. CNT-Confinement Effects on the Menshutkin S_N2 Reaction: The Role of Nonbonded Interactions. *J. Chem. Theory Comput.* **2016**, *12*, 4082–4092.

(110) Tavares, I. S.; Figueiredo, C. F. B. R.; Magalhães, A. L. The Inner Cavity of a Carbon Nanotube as a Chemical Reactor: Effect of Geometry on the Catalysis of a Menshutkin S_N2 Reaction. *J. Phys. Chem. C* **2017**, *121*, 2165–2172.

(111) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

(112) Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **1955**, *77*, 334–338.

(113) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094.

(114) Bakowies, D.; Thiel, W. Semiempirical Treatment of Electrostatic Potentials and Partial Charges in Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Comput. Chem.* **1996**, *17*, 87–108.

(115) Bakowies, D.; Thiel, W. Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Phys. Chem.* **1996**, *100*, 10580–10594.

(116) Huang, X.; Holden, H. M.; Raushel, F. M. Channeling of Substrates and Intermediates in Enzyme-Catalyzed Reactions. *Annu. Rev. Biochem.* **2001**, *70*, 149–180.

6.3 Complementary Information

With all the new implementations needed for **GOCAT** optimization, many of the **GA** operators were newly designed and needed benchmarking in order to create some intuition for this optimization target. This is discussed first in Section 6.3.1. These studies took place before and during the proof-of-concept publication above. In Section 6.3.2, the topic of transferability onto a higher level of theory is addressed, which was also mentioned briefly in the publication but was developed further afterwards. Finally, in Section 6.3.3, typical correlation studies in electrostatic catalysis (reviewed in Section 7.1) are discussed for the Menshutkin reaction. These studies were made during the **DA** reaction computations in Chapter 7 in order to have a comparison with a simple corner case. Needless to say, especially for Sections 6.3.2 and 6.3.3 that concern the Menshutkin reaction again the depictions in the publication are presupposed.

6.3.1 Greedy Benchmark of Meta-Parameters

As **GOCAT** optimization in its current state is a new, unprecedented optimization target, we cannot (easily) refer to other benchmarks or investigations made so far. Indeed, some structural or stackable information is present, *i.e.*, the Cartesian coordinates, in the **GOCATs**. However, using restricted embeddings on which the entities are to be placed as well as not optimizing energies directly, but an agglomerative fitness function with subtle relations, it is not evident *a priori* which **GA** setting should be used in order to deliver enough efficiency and effectiveness. In this context, the “infamous” **no free lunch (NFL)**^[431] theorem(s) should be stated again, following the descriptions in Refs. [29, 155, 156, 450, 451] and interpreted in the context of optimization; there are actually multiple such impossibility theorems also for other contexts: If there exists a set of problems P for which algorithm A prevails over algorithm B by a certain amount, there also exists an equal-sized set of problems P' for which the opposite is true: The **NFL** theorem states that within certain constraints over the range of all possible problems (also theoretically including very extreme ones) every optimization algorithm A will perform as well as every other on *average*, including as the baseline even pure random searches. In other words, there is no general-purpose, *universal* algorithm U that will solve every possible problem P (in finite domains) delivering (near) optimal solutions better than every other algorithm. Rephrasing that more euphemistically, one can, for instance, always incorporate more problem-specific ingredients in order to solve a certain sub-class of problems more efficiently (or effectively) than before, in this trade-off between generality and specificity.² As a corollary of this **NFL**, it is always hard (or even impossible in principle) to find fair comparisons between different algorithms, including not just some (rather subtle) **GA** settings, which is the task of this Section, but more drastic changes as comparing a **GA** with other metaheuristics, such as, *e.g.*, basin-hopping. Thus, one “bare-bone” **GA** implementation (also available in many libraries of

² Note that this also means that plain random search or exhaustive enumeration can be better than a **GA** for certain problems. On the other extreme, one could even think of an algorithm constituted by a simple lookup of an result already known beforehand, which could even run in constant time, though, this would arguably be the most specific “algorithm” possible for delivering back one answer for one specific input.

many programming languages or separate programs) compared to a highly specific and improved version of another algorithm is deemed to be a flawed comparison and should not be generalized.

With these *prolegomena*, the sole purpose of this Section is to find a *rough* guideline for a setting that could be advantageous over others. This was quickly determined in a *greedy* approach. That is, no full meta-parameter optimization of the GA took place, e.g., neither log-grid-based simple enumeration for optimizing meta-parameters as done in ML nor more sophisticated optimizations such as a GA optimizing another GA. Instead, several first optimization runs led to certain inferences that were immediately used in subsequent settings, by sometimes even slightly changing multiple meta-parameters at the same time.³

Moreover, it is *assumed* that these settings are transferable between different GOCAT optimizations, e.g., treating different reactions with unlike reaction paths and other common surfaces or with different model constraints. At least, the moves of the operators within the search space is exactly the same. Yet, whether this translates to the same meaningful moves on the fitness surface when the level of theory or the model constraints are changed, needs further investigations. For other settings such as equidistant charges on a spheres (Chapter 7), there might be different optimal GA settings. For example, for immediate ESP optimization instead of catalysis optimization (Section 6.3.2), the used niching would not be the typical one of the ESP binning described in Section 3.5.3 but an interaction-pair binning one. The latter would be more aware of concrete Cartesian placements and ignorant with regard to the ESP itself since, otherwise, this would completely hinder the ESP optimization.

6.3.1.1 General Setting for the Benchmark

All in all, about 200 different GA configurations were examined by varying the sizes $N_{\text{Ch}} \in \{5, 10, 20\}$ and input settings, some of which will be discussed in the following. As this demands even more computational resources, the benchmark runs were performed on EVB-QMDFE level of theory for a standard Diels–Alder reaction without charge neutrality of the GOCAT, for the same reaction that was also used in Section 3.6.2 (Fig. 3.8 on p. 90).⁴

The general setting common to most of the discussed variants in Table 6.1 on p. 151 is if not explicitly specified otherwise: A pool size of $N_{\text{pool}} = 500$ individuals, ESP vector niching (cf. Section 3.5.3 on p. 83) with at least $N_{\text{deviations}} = 2$ allowed vector element differences (also called deviation count below) bigger than $\Delta\varphi_{\text{ESP}} = 10 \text{ kcal mol}^{-1} \text{e}^{-1}$ at the three stationary frames (R, TS, P). About $N_{\text{GA}} = 7 \cdot 10^5$ GA iterations were performed with a Gaussian

³ Here, we should point to general and different versions of meta-optimization routines in OGOLEM. These might optimize such GA settings *during* the GA itself as a second layer in the future. Certainly, another layer is always more computationally expensive, in general. However, an optimal GA setting as output of such a meta-optimization could then simply be used as input for consequent GA optimizations of the same problem class. What is more, meta-optimizations could bring in *robustness* of these parameters in a more black-box fashion with less burden for the user. A truly adaptive optimization of meta-parameters during the GA could in principle even be better than static pre-settings, as unique per-runtime information could be re-used to guide the adaptation. All this, however, is neither thoroughly tested nor published at the moment.

⁴ Partial charge interaction terms for the energy and the Cartesian gradients were simply added to the usual force field computations of QMDFE. Further, a partial charge interpolation between the QMDFE frames of the reaction path was added, similarly to the energetic EVB coupling to reach the adiabatic PES. The separate EVB-QMDFE potentials for this DA reaction can be seen in Ref. [264, Fig. 2, p. 16717].

rank-based selection operator with a standard deviation of $\sigma = 0.7$ for one individual and thus imposing less fitness “pressure”; the other individual was selected uniformly at random. In the following, mainly the performance of the different mutation and crossover operators as well as their mixture in the GA setting were benchmarked; the usual known dependence on other GA meta-parameters as, for instance, the poolsize, selection pressure, niching strength, etc., are not investigated here, except for some illustrative examples below.⁵ p_{Xover} and p_{mut} , both in the interval $[0, 1]$, describe the crossover and mutation probabilities, respectively: By drawing a uniformly distributed random number x , it is decided, whether to apply one (or maybe also a mixture of) crossover operator(s). If not, a mutation happens automatically instead. Additionally, both can be applied in a row. Thus, we can have “pure” crossover steps ($x \leq p_{\text{Xover}} \wedge x > p_{\text{mut}}$), “pure” mutation steps ($x > p_{\text{Xover}}$) or a mixed step with both applied subsequently ($x \leq p_{\text{Xover}} \wedge x \leq p_{\text{mut}}$).

The main benchmark questions were whether the sweden operator (cf. Fig. 3.3 on p. 79) leads to better performance than the portugal operator (cf. Fig. 3.2 on p. 77), how big and to what mixture mutation operators should be included and whether the canada operator (cf. Fig. 3.4 on p. 81) was a meaningful invention. All the settings discussed here as well as some statistics are compiled in Table 6.1 on p. 151.

Note that the traditional GA algorithm actually was meant to have a high crossover probability (almost $p_{\text{Xover}} = 1.0$) and a smaller mutation probability (maybe about $p_{\text{mut}} \approx 0.05$).^[29] However, one re-occurring observation of GOCAT optimization was that the mutation played a more important role. A possible explanation for this could be that the recombinations used so far are simply introducing bigger jumps through the search space, while the mutations, especially, small step mutations, lead to some “hill climbing” downhill to the next local minimum with the main continuous fitness progression. Moreover, this could also hint at the fact that a very performative crossover operator for GOCAT optimizations has not yet been invented. It should also be stated again (cf. Chapter 3) that we usually do *not* use a local optimization *during* each GA step. This would have to utilize a gradient-free local optimization algorithm due to the complex fitness function. Therefore, we observed clearly a faster fitness progression *without* a local optimization at each step with regard to the absolute wall clock time of the GA run.⁶ Due to this fact, the importance of small mutations in many steps in order to gradually improve on present GOCATs, which is a local exploitation of the information in the population, seems also reasonable.

Hence, the following two different mutations are defined:

- (1) bigMut (big mutation steps): a mixture of 25% re-initialization, 37% charge value, q_i , mutation and and 38% Cartesian moves of the all charges’ positions, r_i . The re-initialization is the random packing of charges with correct constraints directly on the vdW surface, *i.e.*, it is a nullary operator part. Charge values are changed within a Gaussian distribution around the current value, q_i , with a standard deviation of

⁵ Usually, the complete convergence rate can be increased (decreased) by a smaller (bigger) pool size, more (less) selection pressure and less (more) niching and/or diversity checks. Generally speaking, with increasing convergence rate, the possibility to have a prematurely converged solution set rises.

⁶ Each step *with* a local optimization would be worth many *without*, of course; whether the absolute fitness values and/or final GOCATs would differ if a local optimization were used, is yet to be thoroughly investigated.

$\sigma = 0.5 e (q_i \in [-1, +1])$, applied to several charges in one mutation step. Cartesian moves are carried out uniformly at random, up to a maximum of $\|\mathbf{r}_i^{\text{shift}}\| = 5.0$ Bohr for each charge, where $\mathbf{r}_i^{\text{shift}}$ denotes a Cartesian displacement vector.

- (2) smallMut (small mutation steps): a mixture of 25% re-initialization, 35% q_i mutation with a Gaussian standard deviation of $\sigma = 0.1 e (q_i \in [-1, +1])$ around the current q_i for several charges, 40% Cartesian moves up to maximum of $\|\mathbf{r}_i^{\text{shift}}\| = 1.0$ Bohr for each charge.

6.3.1.2 Results and Discussion

An overall impression of the detailed settings is illustrated in Fig. 6.1 on p. 152. These are *averaged* progressions of the *current best GOCATs* known at the respective GA iteration and thus they do not tell anything about all the other individuals in the same population. Two important observations defining the corner cases of this benchmark can be revealed right away:

- The settings that use *only* random-initializations (inp56) or big mutations (even slightly bigger than bigMut, inp55) are the baseline approaches without (a reasonable) information exchange or propagation within the population and these lead to the worst fitness progression and thus severe premature convergence (as expected). They are the cases that any meaningful and effective setting must outperform.
- On the other hand, the canada settings (inp43–inp52) mostly prevail over the others and these are also clearly apparent by the onset iteration number, $N_{\text{canada}} \leq N_{\text{GA}}$, where “kinks” are visible. In these cases, canada was used as a *second* part of a protocol starting with usual mutation steps (like smallMut) and switching to a canada mutation in later GA iterations. The settings that include the canada operators are indicated by the dashed line in Fig. 6.1, *i.e.*, they all end with a fitness below this line.

First block: For a more in-depth discussion, all the settings and their fitness statistics are summarized in Table 6.1 on the next page as well as a more granular comparison is shown in Fig. 6.2 on p. 153. First, the relation between bigMut (*cf.* Item (1) on the previous page) vs. smallMut (*cf.* Item (2)) was tested. These settings start already with a small $p_{\text{Xover}} = 0.2$ in inp19 (as a result of the inputs before inp19, $n < 19$, which are not shown here) and are significantly improved by changing to a smaller mutation operator, smallMut, which can be seen by the drop in the fitness from, *e.g.*, inp19–inp22 to inp23 and the subsequent inputs. In inp22, also a smallMut was used, but at the same time more than one meta-parameter was changed, the $p_{\text{Xover}} = 0.8$. Thus, this run is worse by not harnessing the (more effective) small mutation, *i.e.*, including this only as a very small fraction in the operator mixture. Next, the more important observation was made between inp23 to inp35 that the sweden operator is beneficial with regard to the fitness progression (*cf.* inp25) as now fitness values around 36–40 are reached. This was tested multiple times with smallMut as a mutation and with $p_{\text{Xover}} \in \{0.2, 0.5, 0.8\}$. Following the lead of not introducing potentially additional

Table 6.1: Descriptions of the 39 GA settings shown in Figs. 6.1 and 6.2: The mean final best fitness reached after the all GA steps, \bar{f} , its standard error, σ_{dev} , in parentheses as well as the minimum and maximum fitness interval at the end, $[f_{\text{min}}, f_{\text{max}}]$, are compiled. Further details are given in the main text, where each of the four blocks is separately discussed.

input setting	description	$\bar{f}(\sigma_{\text{dev}})$	$[f_{\text{min}}, f_{\text{max}}]$
inp19	portugal with 1 and 2 cuts ($p_{\text{Xover}} = 0.2, p_{\text{mut}} = 0.1$), bigMut	46.0(2.3)	[42.0, 51.0]
inp20	inp19, but with 4 and 3 cuts	45.5(1.9)	[40.2, 48.7]
inp21	inp20, but $p_{\text{Xover}} = p_{\text{mut}} = 0.5$	45.9(2.0)	[41.5, 50.9]
inp22	inp19 with $p_{\text{Xover}} = 0.8, p_{\text{mut}} = 0.1$ and smallMut	44.2(3.3)	[38.2, 50.4]
inp23	sweden cuts, $p_{\text{Xover}} = 0.2$, 2 different planes, smallMut	38.1(3.3)	[32.8, 46.4]
inp24	sweden cuts, $p_{\text{Xover}} = 0.2$, 1 common plane, smallMut	39.7(5.9)	[32.1, 61.1]
inp25	inp23, but with $p_{\text{Xover}} = 0.8$	38.7(1.8)	[34.8, 41.4]
inp26	inp24, but with $p_{\text{Xover}} = 0.8$	40.3(3.7)	[33.7, 47.7]
inp27	portugal (1 & 2 cuts to $p = 0.2$ & 0.2), sweden (both variants to $p = 0.3$ & 0.3), $p_{\text{Xover}} = 0.8$	37.0(2.5)	[32.7, 42.7]
inp28	inp27, but with $p = 0.1$ & 0.1 for portugal and $p = 0.4$ & 0.4 for sweden	36.6(2.3)	[33.1, 42.1]
inp29	inp28, but with $p_{\text{Xover}} = 0.2$	36.4(3.9)	[32.5, 52.3]
inp30	inp28's crossover, bigMut again ($p_{\text{Xover}} = p_{\text{mut}} = 0.5$)	42.5(1.7)	[38.3, 45.2]
inp31	inp27's crossover, ($p_{\text{Xover}} = 0.5, p_{\text{mut}} = 0.1$), smallMut	36.5(2.5)	[32.7, 41.7]
inp32	inp27's crossover, bigMut, ($p_{\text{Xover}} = 0.5, p_{\text{mut}} = 0.1$)	40.3(2.4)	[36.0, 44.2]
inp33	inp27's crossover, bigMut, ($p_{\text{Xover}} = 0.2, p_{\text{mut}} = 0.1$)	42.4(1.6)	[39.1, 44.9]
inp34	inp27's crossover, bigMut, ($p_{\text{Xover}} = 0.8, p_{\text{mut}} = 0.1$)	41.1(3.0)	[34.6, 49.0]
inp35	inp28's crossover, bigMut, ($p_{\text{Xover}} = 0.8, p_{\text{mut}} = 0.1$)	40.6(2.0)	[37.2, 44.5]
inp36	inp31, maximum of 5 GOCATs per niche	36.6(2.8)	[33.4, 45.0]
inp37	inp31, maximum of 10 GOCATs per niche	37.5(3.3)	[32.9, 45.3]
inp38	inp31, maximum of 2 GOCATs per niche, $\Delta\varphi_{\text{thresh}}^{\text{ESP}} = 20 \text{ kcal mol}^{-1} \text{e}^{-1}$	36.3(1.6)	[33.5, 40.6]
inp39	inp31, maximum of 2 GOCATs per niche, $\Delta\varphi_{\text{thresh}}^{\text{ESP}} = 5 \text{ kcal mol}^{-1} \text{e}^{-1}$	38.4(3.0)	[32.8, 45.0]
inp40	inp31, maximum of 2 GOCATs per niche, $N_{\text{deviations}} = 6, \Delta\varphi_{\text{thresh}}^{\text{ESP}} = 10$	35.8(2.0)	[32.8, 41.3]
inp41	inp31, $N_{\text{pool}} = 2000$, maximum of 8 GOCATs per niche, $\Delta\varphi_{\text{thresh}}^{\text{ESP}} = 10$, deviation count of 2 again	38.1(1.8)	[35.4, 41.7]
inp42	inp31, no niching at all	41.7(4.3)	[33.2, 51.0]
inp43	inp31, mixed giant step canada (for $N_{\text{GA}} \geq 630 \cdot 10^3 = N_{\text{canada}}$), max. of 3 GOCATs per niche, $\Delta\varphi_{\text{thresh}}^{\text{ESP}} = 10, N_{\text{deviations}} = 2$	36.5(2.5)	[32.9, 42.1]
inp44	inp43 with mixed giant step canada (for $N_{\text{GA}} \geq 350 \cdot 10^3 = N_{\text{canada}}$); selection pressure $\sigma = 0.1$	38.8(2.8)	[33.3, 44.0]
inp45	inp43 with pure small step canada (for $N_{\text{GA}} \geq 490 \cdot 10^3 = N_{\text{canada}}$)	34.0(3.7)	[29.9, 45.2]
inp46	exact inp31 with pure small step canada (for $N_{\text{GA}} \geq 490 \cdot 10^3 = N_{\text{canada}}$)	33.6(3.4)	[29.6, 44.9]
inp47	inp45, but with mixed small step canada	34.8(3.9)	[29.9, 45.9]
inp48	inp45, with pure small step canada (for $N_{\text{GA}} \geq 140 \cdot 10^3 = N_{\text{canada}}$)	32.5(2.9)	[28.8, 41.1]
inp49	inp45, with mixed small step canada (for $N_{\text{GA}} \geq 140 \cdot 10^3 = N_{\text{canada}}$)	33.0(3.3)	[28.4, 44.8]
inp50	inp48, pure canada from the beginning	30.6(1.5)	[28.2, 34.4]
inp51	inp49, mixed canada from the beginning	30.6(1.7)	[28.4, 34.8]
inp52	inp51, multi-mixed canada from the beginning (<i>i.e.</i> , small step, giant step canada, and q_i and Cartesian mutation)	31.5(2.3)	[28.1, 38.1]
inp57	inp52 with $N_{\text{GA}} = 2 \cdot 10^6, N_{\text{pool}} = 600, \Delta\varphi_{\text{thresh}}^{\text{ESP}} = 20 \text{ kcal mol}^{-1} \text{e}^{-1}$	30.0(0.7)	[28.9, 32.0]
inp53	baseline: small mutations (of inp31) only ($p_{\text{Xover}} = 0.0$)	44.3(4.6)	[36.8, 54.7]
inp54	baseline: multi-mixture of mutations (small and big ones, $p_{\text{Xover}} = 0.0$)	45.5(3.4)	[39.6, 51.9]
inp55	baseline: just random initialization (no hill climbing, no pool-information propagation ($p_{\text{Xover}} = 0.0$))	131.0(11.6)	[111.3, 153.7]
inp56	baseline: big mutations only ($p_{\text{Xover}} = 0.0$)	72.8(2.8)	[66.9, 79.9]

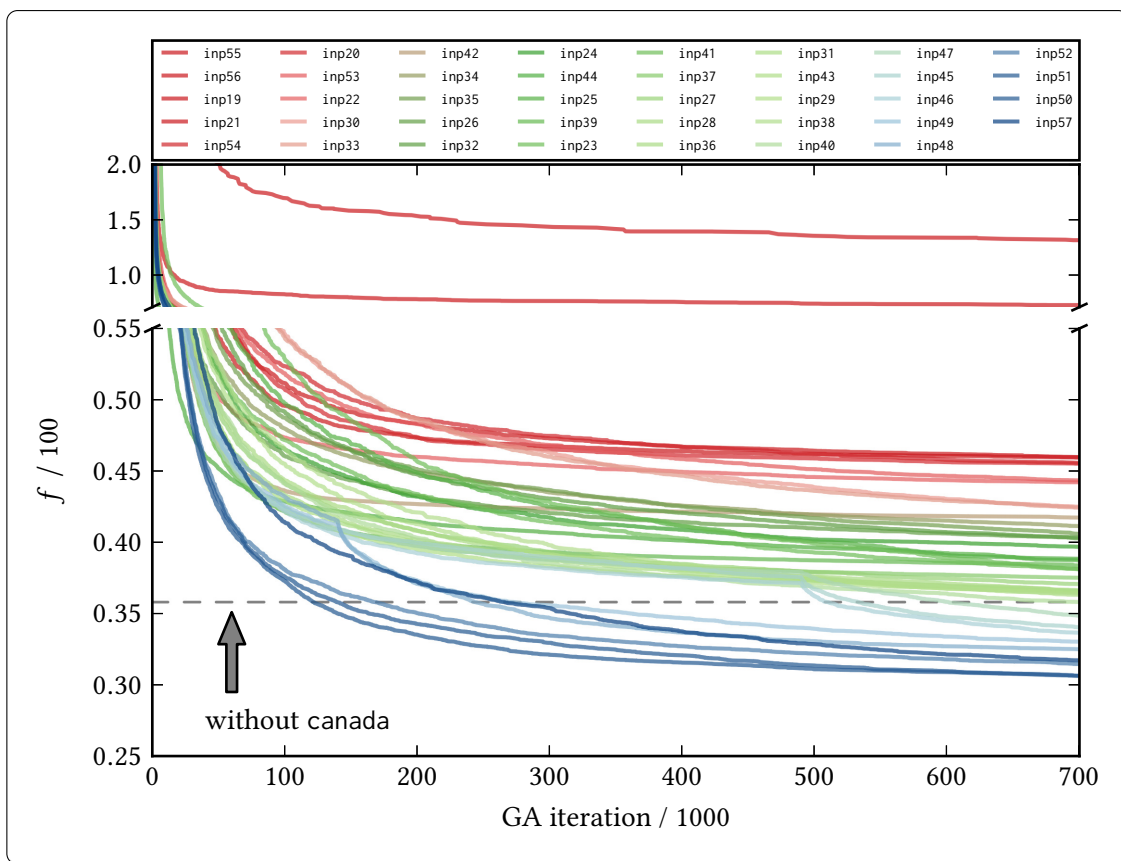


Fig. 6.1: Each *current best rank* of 31 different GA runs are *averaged* for the fitness progressions shown here; this is done for each operator setting separately, inp_n , of which there are 39 variants shown ($n \in [19, 57]$). The color map (from red over green to blue) is chosen based on the final fitness at the end and the legend labels are sorted accordingly, starting with inp_{55} as the worst setting and ending with inp_{57} as the best one. The dashed line divides settings that did not use *canada* (above this line) from those settings that used this operator (below this line). For details about each input setting, see Table 6.1 on the preceding page and the main text. Part I (more details below in part II, Fig. 6.2).

bias into the combination of operators, a mixture of multiple portugal cuts and sweden cuts were also used, as, *e.g.*, in inp_{27} and the following, which gave a (tiny) performance benefit. Their probabilities in the mixture of using the sub-variants of the crossover operators, *i.e.*, number of cuts for portugal or orientation of the 2D plane for sweden, is denoted with a p in Table 6.1. As a first conclusion, inp_{31} constitutes one general setting that was used for many optimizations, including the one of this Chapter and the publication. This setting consists of the `smallMut` and a mixture of both portugal (two different cut versions) and sweden in two variants (*cf.* Fig. 3.3 on p. 79), with a $p_{\text{Xover}} = 0.5$ and $p_{\text{mut}} = 0.1$.

Second block: In the next part in Table 6.1, some (more or less educative) niching benchmarks were done. Generally, in *GOCAT* optimization we have never observed a final population with insufficient intrinsic coordinate variability in the final pool as long as *some* niching is used. In contrast, in Chapter 5 for *CSO*, we usually know the final best fitness (*i.e.*, energy) of the global minimum of those benchmark problems that are already well

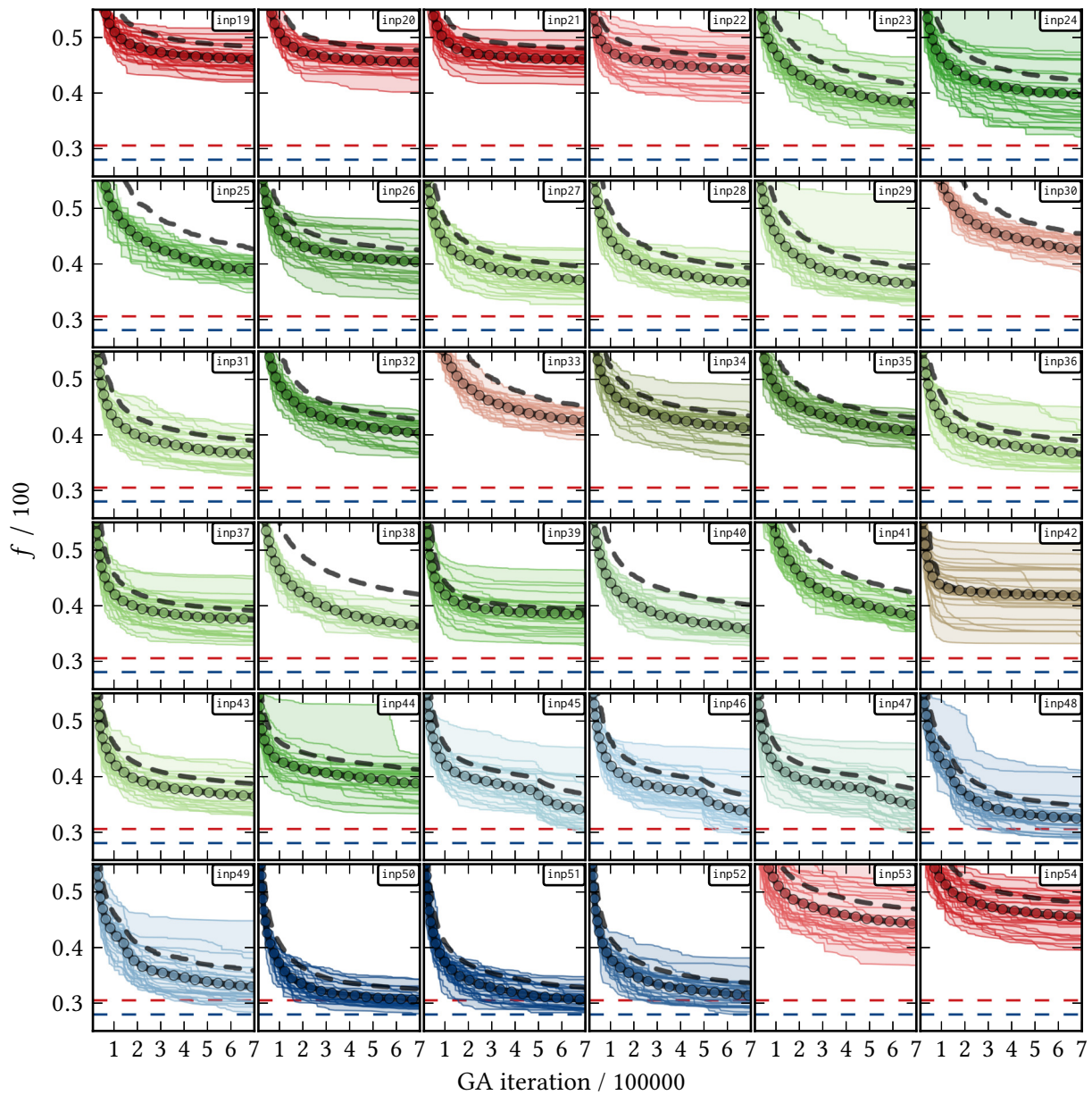


Fig. 6.2: Complementary to Fig. 6.1 on the preceding page, 36 (of the 39) settings are illustrated separately here; the worst baseline approaches, inp55–inp57, are missing. The *averaged* fitness progressions of *the best ranks* of 31 separate GA runs (the same as in Fig. 6.1, also using the same colormap but a contextual sorting) are shown with line-dots here, while *each separate* run that is used for the averaging is plotted in the back. Dashed gray lines show the *mean* of the average fitness of the *full* population, *i.e.*, by taking into account all GOCATs, not just the current best ranks. The red dashed line at $f = 30.6$ denotes the lowest *averaged* fitness reached and the blue dashed line at $f = 28.1$ the minimum fitness ever occurred. For details, see Table 6.1 on p. 151 and the main text. Part II.

investigated. This minimum is often hard to detect unless enough intrinsic (genotypical) information is conserved by niching in order to not converge too early within deceptive basins of attraction that cannot easily be left by a converged GA anymore. Conversely, in GOCAT optimization often many individuals which are comparably fit are found in the final pool that do not share many (obvious) similarities, at least *prima facie* by looking at Cartesian coordinates, for instance. Rather, very unlike GOCATs do end up in the final pool by using some niching, while the fitness and reaction barriers are very similar. Thus, in many cases we observed multiple “global optima” on the fitness surface, or at least competitive local optima without a clear-cut chemical distinction or domination of specific solutions. As a result, we cannot ask a similar question (yet), which exact niching in what configuration might lead to the best (chemical) GOCATs.⁷

In order to get an impression of the mentioned convergence behavior, the niching strength was varied, e.g., by increasing the allowed *maximum number of GOCATs in one niche* (inp31, inp36 to inp37) or by changing the allowed difference of $\Delta\varphi_{\text{thresh}}^{\text{ESP}}$ on the same atom to be treated as being similar (inp38, inp39). If $\Delta\varphi_J^{\text{ESP}} > \Delta\varphi_{\text{thresh}}^{\text{ESP}}$ on the same atom, J , of two GOCATs in the comparison (*cf.* the niching depictions in Section 3.5.3), the deviation count is incremented; this maximum allowed deviation count, $N_{\text{deviations}}$, to identify two individuals as being similar was changed in inp40. Additionally, a very large pool with a larger niche population was also tested (inp41) and, as a baseline, no niching at all (inp42). Hence, some intuitive and expected tendencies can be shown as, for instance: On the one hand, the convergence slows down significantly in inp38 (more severe niching, similarly in inp40), especially apparent by the gray dashed line as the mean fitness of the total populations and its bigger shift from the best fitness (dots) than in most other settings. Note that in these cases, and even more so in inp41 with the giant pool, the runs are not converged sufficiently, *i.e.*, many more iterations would have been needed to reach the final best fitness.⁸ On the other hand, the convergence rate increases, which is observed in inp38 or inp42 with a weak or *no* niching at all, respectively. Note the completely overlain mean fitness and mean best fitness in the case of no niching, which points to the fact that each individual in each pool is essentially (almost) equal. Moreover, the spreads of the separate runs of the 31 are very different (background), *i.e.*, the GA completely focuses the pool into separate basins of attraction varying in each of the 31 runs and cannot jump out of those anymore.

Third block: The next part in Table 6.1 with the settings inp43–inp52 and inp57 (the last one was calculated afterwards as convergence check) investigated the performance of the canada mutation operator. Because this was meant to be an “end-game” operator learning from the emerged chemically relevant ESP patterns and sampling new GOCATs

⁷ Note, however, that such a specific question is not asked in Chapter 5 in this form as well. Rather, we generalized this view into the more general need of having an order parameter during the GA and not by asking which one in particular. An answer to this latter question might also be already too confined (*cf.* the NFL).

⁸ As the settings became more effective, more total iterations would have been needed in some cases. Instead of calculating everything anew, some benchmarks with a significantly increased iteration number were done in order to reach a fully converged pool and a final best fitness (*vide infra*).

based on these, a schedule of operators was applied that changes the mixture of operators during the GA and is based on the iteration number. First, a usual setting such as one of the aforementioned is used and then, after some progression, the canada mutation, often also in a mixture with other operators, is switched in at the iteration N_{canada} and used in all subsequent iterations.

Here, we can simply state that canada brings in a severe performance benefit if it is used with a smallMut setting rather than with a bigMut setting (*cf.* the depictions for canada in Fig. 3.4 on p. 81); this is apparent when comparing inp43, inp44 with the others from this block. With a big mutation setting, the next reached ESP-based local optimum has a bigger deviation from the target ESP and this might change the GOCAT quite severely such that the acceptance rate becomes quite small. This is also caused by the gradient problems that are discussed in a different context in Section 6.3.2. Furthermore and surprisingly, canada is even better when it is used directly from the beginning, contrary to what was intended before (inp50, inp51). Even a pure canada, *i.e.*, as the sole mutation operator without any others, is well-performing; the mixed version is one that includes also the usual Cartesian and charge mutations as smallMut. The most general setting, actually, would be the inp52 which is constituted by 5% re-initialization, 15% charge mutation, 15% Cartesian mutation (which share the same meta-parameters as smallMut defined above, see Item (2) on p. 150), with now additional 50% small step canada (wrapping smallMut, except for the re-initialization part) and 15% big step canada (wrapping bigMut, see Item (1) on p. 149). This is denoted as “multi-mixture”. It might bring in the least artificial bias since even some usually less effective, but more systematically explorative moves are included. This is combined with the mixture of portugal and sweden of inp31 as crossover operators. This setting was also used for $N_{\text{GA}} = 2 \cdot 10^6$ iterations again in inp57 to reach an even better converged pool, see Table 6.1. In this case, just a small overall improvement is seen approaching final mean best fitness values of $\bar{f} \approx 30.0$, which depicts a good estimate of the minimum fitness that can be reached with $N_{\text{Ch}} = 10$ (and the other GOCAT model parameters). This can also serve as boundary for the not-yet converged results, as, *e.g.*, in inp38 or inp41.

Fourth block: The last part in Table 6.1 contains the baseline approaches: inp53 is *only* a smallMut setting that already performs quite acceptably except for the wide spread which is to be expected; this is a population-based hill climbing (*cf.* Algorithm 2.2 on p. 26), essentially. inp54 is significantly worse using just bigMut. This restates that small-step mutations are needed for gradual exploitation of the basins of attraction and are more effective than the big steps. In inp54, multiple different mutations are used, *i.e.*, small, intermediate and combined ones, similar but not significantly different from plain smallMut of inp53. inp55 is the raw baseline approach with *just* re-initializations directly on the vdW surface of new GOCATs, *i.e.*, using just a nullary operator without any information exploitation as a fully explorative setting and is by no means competitive to the GA settings above.

Further discussion: Lastly, some reservations of the discussed results could be expressed. Indeed, what is not included above, since just the fitness is represented, is the actual intrinsic variability of $\{q_i, \mathbf{r}_i\}_{i=1}^{N_{\text{Ch}}}$ of the **GOCATs**, *i.e.*, the actual genotype diversity. Certainly, the difference between the best and the mean fitness can be used (and was used above) as a proxy for the conserved diversity, but more in-depth investigations are definitely needed to answer questions as, for instance, which setting (and which niching) might be best to reach chemically meaningful **GOCATs**. This is amplified by the fact that the final fitness itself is poorly discriminating quite unlike **GOCATs** as pointed out above when speaking about multiple similar local optima without clear dominance. Besides, some of the settings lead to fitness progressions that are not so different from the others, at least not if using a full inferential statistical treatment. That is, some of the final results of the settings would not be significantly different from each other (within the error margins). As a result, no definitive statements (with a vanishing type one error of inferential statistics) about the performance order could be made; many settings result in a mean fitness in a small interval, $\bar{f} \in [40, 44]$. However, the main observations *are* significant, as they are: smallMut is better than bigMut, sweden is needed and canada brings in the main improvement.⁹

In Appendix A.1 on p. 265, complementary benchmark studies with $N_{\text{Ch}} \in \{5, 20\}$ are given. Two further insights from these can be summarized as follows: (1) Small step mutations almost suffice to optimize a small $N_{\text{Ch}} = 5$ model. In this case, the problem size and search space is small enough to use a population-based hill climbing. In contrast, for $N_{\text{Ch}} = 20$ crossover steps dominate the progression efficiency such that sweden plays an even more important role. At the same time, canada is beneficial, but this is not as apparent (*i.e.*, without the clear onset impact) as in the $N_{\text{Ch}} = 10$ case discussed above. (2) As expected, any increase of model complexity leads to a smaller final fitness and thus to better catalytic reaction profiles. Though, any possible translation to a real molecular embedding or simple interpretation of the results is hampered.

Possible Future Extension of Operators: More generally, canada could be used when generalized also as a wrapper for arbitrary other crossover and/or mutation operators. For example, also after the sweden phenotype cut planes with more conservation of local structure than after arbitrary displacements of a mutation, a canada step could be used to optimize the **GOCAT** with respect to the **ESP**. This could also be tried as a linear regression problem by only optimizing the charge values, $\{q_i\}$, (*cf.* Section 2.1.1.2), after sweden changed the Cartesian coordinates. Other hybridization protocols might also be interesting—*i.e.*, more complex steps that already conserve some phenotype character such as the local structure in the **GOCAT**—by incorporating, *e.g.*, “local heat pulses”.^[406]

⁹ Actually, the current author has never experienced such elaborated comparisons and statistical significance tests in the context of global optimization in chemistry. However, as stated in Ref. [156], one should even use *independent* runs for each test combination when intending to compare more than just two settings (otherwise a *t*-test might be sufficient), *i.e.*, about at least 30 runs for inp19 vs. inp20, 30 *new* runs for comparing inp19 vs. inp21, etc., and using something like **analysis of variance (ANOVA)**. Here, significantly more computational resources would be needed to follow such a statistically rigorous and solid approach due to the scaling of $O(n_{\text{settings}}(n_{\text{settings}} - 1) \cdot N_{\text{runs}})$, which leads to 45,942 instead of the 1209 separate **GA** runs shown in this Section, and each should also have more total N_{GA} iterations.

Additionally, some simplification operators could also be introduced that explicitly search for symmetric representations by using something similar as `canada` (that implements all of Section 2.1.1.2) with symmetry constraints to impose the same **ESP** at symmetrically exchangeable atoms.

6.3.2 Transferability by Translation Protocols

As global optimization in general is very computationally expensive—at least if a global convergence is intended and not just some “short-cuts” to some well-enough **GOCATs** due to the first more than exponential fitness decrease after start up—a compromise with regard to the level of theory is always needed. For instance, just *one* setting in the publication for the $N_{\text{Ch}} = 10$ case with one single fitness function and charge neutrality (Section 3.3 in the publication starting on p. 137), consumed the following number of total **SP** evaluations, N_{SPs} , where each single point can be an energy and/or a Cartesian gradient evaluation of one frame within a **GOCAT**. With $N_{\text{GA}} = 1.2 \cdot 10^6$ **GA** iterations, two children created and evaluated in each **GA** iteration, a pool size of $N_{\text{pool}} = 600$, a frame number of $N_{\text{frames}} = 21$ of the **MEP** and 10 separate (farming) **GA** runs, we have

$$N_{\text{SPs}} \leq \underbrace{\left[\left((1.20 \cdot 2) \cdot 10^6 + 600 \right) \cdot 21 \right]}_{\text{total GOCATs per run}} \cdot 10 + \underbrace{\left(4.02 \cdot 10^6 \cdot 21 \right)}_{\text{total SPs of loc. opt.}} = 588 \cdot 10^6, \quad (6.1)$$

where the second term denotes the complete local optimization of the accumulated pool having now $N_{\text{GOCAT}} = 6000$ **GOCATs** for this specific case. The local optimization usually is applied *once* after the separate **GA** runs, not during or within each **GA** iteration (see Chapter 3). The calculated number is an upper bound for this specific setting since the `immediateFallback` (which was explained around Algorithm 3.1 on p. 87) will save a few **SPs** in each run—very dependent on the overall unique progression of the particular optimization—but the order of magnitude will be correct.¹⁰ That is, this single entire pool of 6000 individuals at the end of the merged 10 pools of $N_{\text{pool}} = 600$ each needed about half a billion **SPs**. Having to calculate one chemical reaction with different settings, *e.g.*, with changing N_{Ch} , fitness functions, restraints on the **GOCAT** model, etc., usually about several of those total runs of Eq. (6.1) are performed. From this perspective, it is simply not computationally feasible to compute all this on a higher level of theory, most probably on a **DFT** level, unless utilizing very extended **HPC** computational resources that are more extensive than available on a daily basis.

6.3.2.1 Protocols for Translation

Hence, already in the publication of this Chapter, we wanted to *translate* between levels of theory trying to re-use information we have gained by optimizing on the **SQC** level. The main problem is that the **MEP** treated on varying levels will be (at least) slightly different,

¹⁰ One evaluated number for this same setting is $N_{\text{immediateFallback}} = 10.04 \cdot 10^6$ function calls with a total of $192.54 \cdot 10^6$ saved **SPs**. Consequently, still approximately $400 \cdot 10^6$ **SPs** are needed for one **GOCAT** database.

which is also already the case in the simple Menshutkin reaction. In particular for the this reaction treated on PBE0/def2-TZVP^[201,452] vs. PM7,^[206] the attacking NH₃ has a longer distance to the attacked C atom ($d = 3.33 \text{ \AA}$ vs. $d = 3.03 \text{ \AA}$ on PM7) in the loosely associated adduct—which we call **R** throughout this Thesis. The same is true for the distances between C and Cl in the contact ion pair, the **P** frame ($d = 2.70 \text{ \AA}$ vs. $d = 2.59 \text{ \AA}$ on PM7), while the **TS** is already more product-like on DFT. These small Cartesian deviations for the **MEP** that are reflected in the different discrete **NEB** frames will result in a (slightly) variant **vdW** surface such that a **PM7 GOCAT** cannot simply be put around the same reaction on another level without subsequent translation protocols.¹¹ In this regard, two fundamentally different translation algorithms were implemented that will be compared in four “flavors” below:

- (1) New Cartesian charge positions on the higher level of theory are optimized with respect to the full distance matrix with the elements $D_{iJ} = 1/r_{iJ}$. In this case, each distance, r_{iJ} from a charge i to an inner atom J is calculated in the starting individual (**PM7**), which depicts the reference distances, and the target is to reach those same reference distances again when surrounding the new structures (**DFT**). Thus, no inter-charge distances are included. This optimization problem usually generates an over-determined system and is solved in a least-squares sense with equal weights on each distance, by optimizing the Cartesian charge positions only. (Indeed, simple translations and rotations can be re-mapped perfectly, of course, but with changes of the exposed **vdW** surface, this translation is not expected to result in zero for the objective function).¹² Usually, only the three stationary frames (**R**, **TS**, **P**) of the source and target **MEP** are used since the frame number between both **NEBs** can also vary. As the charges are now placed in a way that they approach the source distances as close as possible, they will not lie on the **vdW** surface. Hence, as a next step the **vdW** remapping takes place—which then partially destroys the best positions found by the first step.
- (2) Instead of employing merely the Cartesian distance matrix, a (partial) Coulomb matrix $C_{iJ} = q_i/r_{iJ}$ with charge values q_i in the numerator is utilized in this second protocol. The first protocol was exactly the one already used in the publication of this Chapter, whereas the second one was not invented back then. The objective function that is used in this protocol is based on the **ESP** values at the atoms (or more specifically the **ESP**-differences); this function was explained in Section 2.1.1.2 for Eq. (2.14) on p. 18. It serves for subsequent local and/or *global* optimizations. Therefore, for the Menshutkin reaction, a C_{3v} -symmetrized **ESP** optimization can take place by using our **GA** again but this time fully hybridized with a local optimization in each iteration to reach the (globally) best **GOCATs** with the same scalar **ESP** on a subset of positions, *i.e.*, at selected frames’ atoms (**R**, **TS**, **P**).

In the following, the two methods are compared in four different settings: The protocol

¹¹In the simplest version, these at least have to include some translation/rotation and some **vdW** remapping.

¹²Typically, this will also only result in a local optimum.

that is named “vdW”¹³ is a simple remapping onto the new vdW surface, without any distance or Coulomb interaction optimization. It is the simple mapping that always occurs in each iteration during the GA, *i.e.*, it is used then to satisfy the usual restraints and not to translate between different vdW surfaces as in the present case. The protocol “distances” terms Item (1) above, the protocol “ESP glob.” the *global* optimization using the method of Item (2), and the protocol “ESP loc.” is the *local* version of the same method of Item (2). For all translations, about 250–300 HC clusters were used from an accumulated population of the corresponding GOCAT setting. Then, the best rank as well as the nearest neighbor to the calculated cluster mean in a 27 dimensional ESP space were used as starting individuals, as long as the best and the nearest neighbor GOCATs are not the same (by chance)—thus the irregular observation number, N_{GOCAT} , follows below; this is similar to the methodology of the publication of this Chapter. Furthermore, a standard pure random initialization is carried out that does not capitalize any information acquired from the SQC pool beforehand, *i.e.*, the nullary initialization directly on the DFT level of theory. This is the “random” protocol.

6.3.2.2 Results and Discussion

A first translation is shown in Fig. 6.3 on the following page for $N_{\text{Ch}} = 3$ systems *without* total charge neutrality, $\sum_i q_i \neq 0$. The threshold fitness value that should not be surpassed is about $f_{\text{thresh}} = 409.1$ on DFT level of theory. This is solely a fitness value based on the barrier itself, while all other terms in Algorithm 3.1 on p. 87 will be zero in the gas phase path. Obviously, all protocols detailed above can optimize many GOCATs that already are in the catalytic window, *i.e.*, that are showing smaller fitness values than the pristine gas phase path. In the random initialization, one single GOCAT is near the threshold (at about $f = 450$), but most of these are much worse (note the logarithmic scale in the left panel of Fig. 6.3).

On the left-hand side of Fig. 6.3, there is a distinct bunch of points scattering around values of about 1400 and 2400 fitness points. This is due to the *discrete* fitness penalty (coming in chunks of 1000 points in this case) if the frame indices of the stationary points for {R, TS, P} are shifted. Often, a slight overstabilization at frame index > 1 with energies smaller than the actual R energy (at frame 0) with respect to any last decimal places can lead to fitness values that should also lie at around 350–400, but do so at 1350–1400 with the penalty. This penalty is most pronounced in the ESP optimization method. Presumably, these can be explained due to overfitting as the ESP optimized GOCATs are different with respect to all intrinsic coordinates, $\{q_i, \mathbf{r}_i\}$, both Cartesian coordinates and charge values, with possible bigger changes from the starting candidate solution. This is already true in the locally optimized method, and even more so when the full-blown global optimization is used, whereas the distance optimization was always *locally* and only treated the Cartesian coordinates.¹⁴

¹³Note that when quotation marks are used in the main text in the following, these four translation methods are directly addressed.

¹⁴Generally, in the local optimization of the fitness after a full GA, it was observed that right before the actual convergence of, *e.g.*, BOBYQA, the fitness values were decreasing continuously, but then led to erratic sudden jumps. This points to the fact that the local optimization progression of each GOCAT can already result in

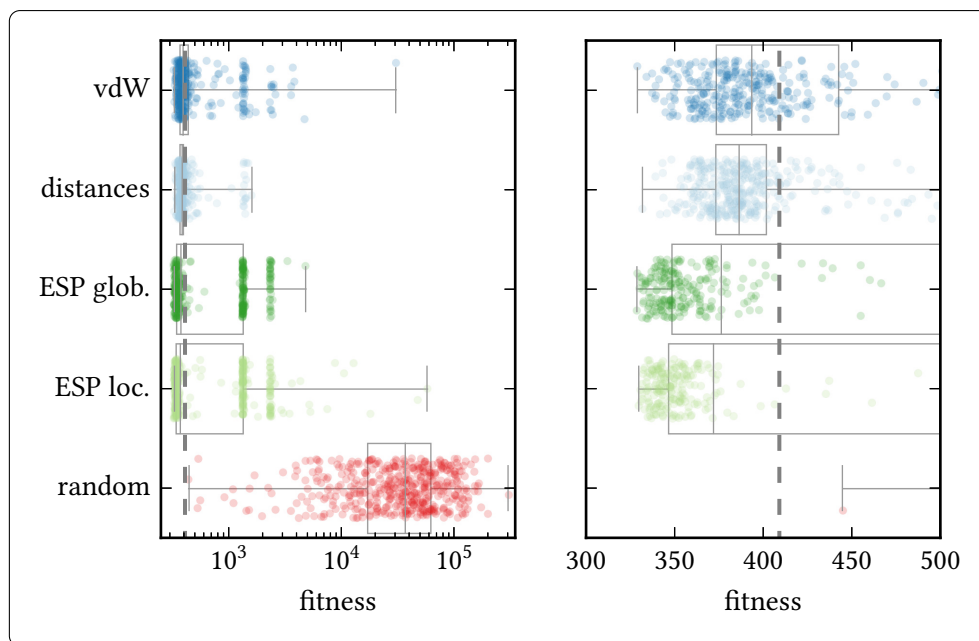


Fig. 6.3: $N_{\text{Ch}} = 3$ (non-neutral): Fitness values after translation from PM7 to PBE0/def2-TZVP level or theory using the four different variants (explained in the main text around Items (1) and (2) on p. 158) and compared to the baseline approach, the *random* initialization, plotted in two different regions/scales. The y -axis (ordinate) of this plot is categorical with added random jitter/spread to see the separate data points, and with a boxplot for quartiles and the median value in the background. The vertical dashed line is the fitness of a corresponding pristine gas phase Menshutkin reaction on this DFT level without a GOCAT with $f_{\text{thresh}} = 409.1$ (the statistics are given below in Table 6.2 on p. 164). Values with $f < f_{\text{thresh}}$ can then be considered as being catalytic.

The corresponding energies as well as Cartesian energy gradient norms are shown in Fig. 6.4 on the next page. Here, the observations solidify that the energies are very similar in both protocols and show the lower barrier (except for some outliers). The gradient norms are due to the gradient norm penalties in a region of about $\|\nabla E\| \approx 10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ (compare also with Table 6.3 on p. 165). Contrariwise in the random case, a typical variance of energies and gradient norms can be observed.

Note that many of the randomly initialized GOCATs lead to not converging SCFs at least at one frame; in particular, this happened in 539 cases here (*cf.* Table 6.2 on p. 164). The fitness values that would just be an extremely high offset are not shown here. This is expected as, because of the completely random initialization, many extreme, chemically meaningless surroundings are generated that cannot even be evaluated at all due to convergence problems of the QM calculations.

A harder translation problem for the same Menshutkin reaction onto the DFT level is similarly shown in Fig. 6.5 on p. 162 for the fitness values and Fig. 6.6 on p. 163 for energies and gradient norms of the GOCATs for the problem size of $N_{\text{Ch}} = 10$ and for neutral surroundings. The anticipated tendency here is that the random initialization has

a “bouncing” against the discrete penalty walls right before having found the local optimum such that the returned optimum is the smallest fitness in the surrounding capped by the penalty wall. Now, it is assumed that a “perfect” representation of a reference ESP with the ESP of the new path also leads to similar artifacts.

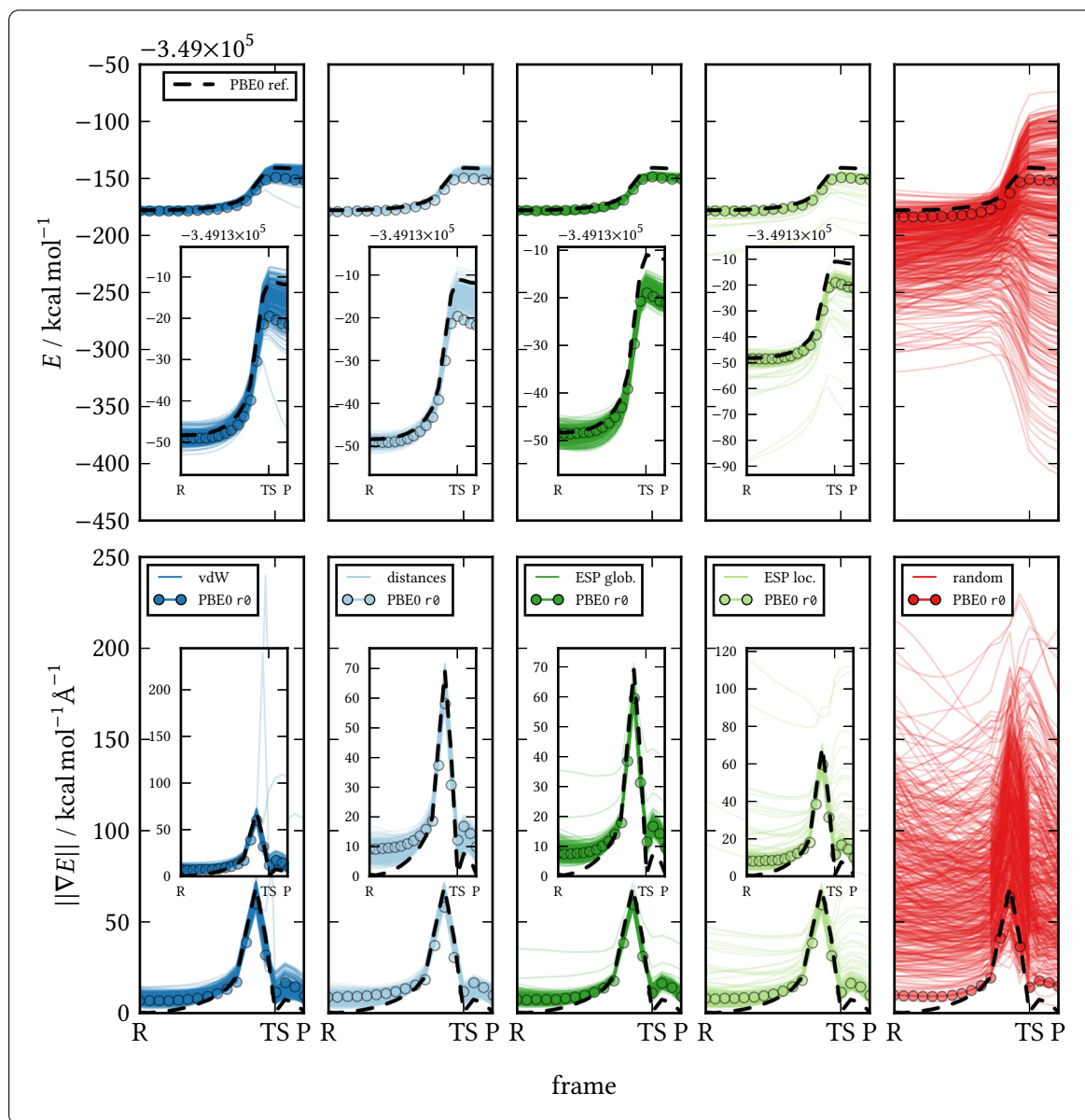


Fig. 6.4: $N_{\text{Ch}} = 3$ (non-neutral): Corresponding reaction energy profiles and energy gradient norms for all data in Fig. 6.3 for the four translation protocols and the random initialization. Corresponding statistics are given in Table 6.3 on p. 165. “PBE0 r_0 ” is the case-specific best GOCAT with the lowest fitness *after* the translation, which is not the r_0 on PM7 before the translation.

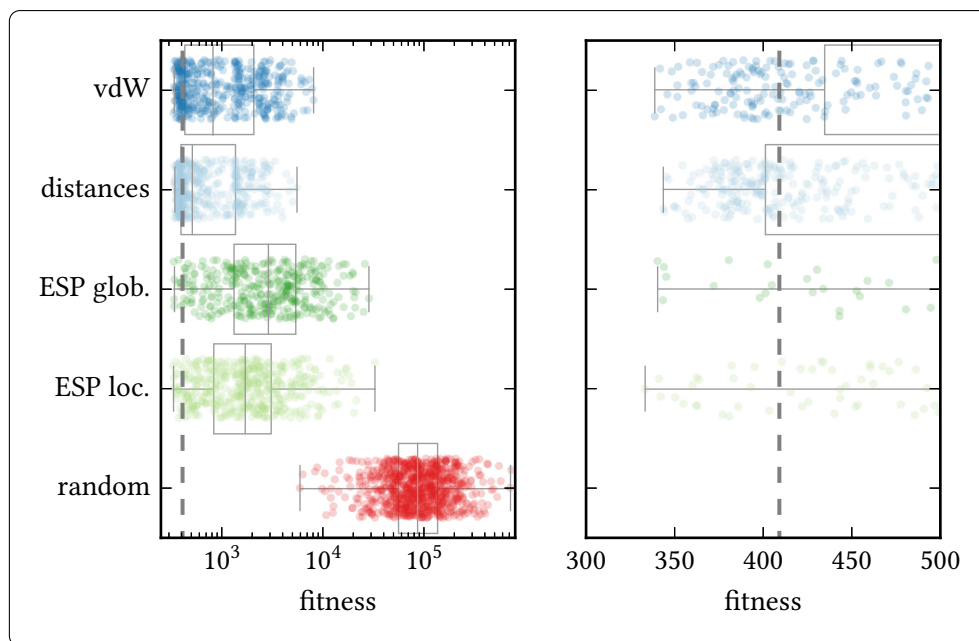


Fig. 6.5: $N_{\text{Ch}} = 10$ (neutral): Fitness values after translation from PM7 to PBE0/def2-TZVP for the four translation protocols and the random initialization. For plotting details see Fig. 6.3 on p. 160.

even greater problems in finding some promising starting **GOCATs** as the search space increases. At the same time, the *neutrality* constraint, $\sum_i q_i = 0$, of each **GOCAT** already filters out some very extreme electrostatic surroundings such that actually more than in the $N_{\text{Ch}} = 3$ case can be converged. Admittedly, the simpler distance-based least squares optimization is quite more efficient in finding already catalytic **GOCATs** in contrast to the more elaborated optimization of the full **ESP**, both locally and globally.

Some corresponding descriptive statistics is compiled in Table 6.2 on p. 164. For the $N_{\text{Ch}} = 3$ case, we have about 63.8%, 80.2%, 56.2% and 52.9% of **GOCATs** for the four different variants with fitness values below 409.1 that could be considered as already having *some* catalytic effect onto the reaction. For $N_{\text{Ch}} = 10$, these relations decrease severely to 18.6%, 29.7%, 2.8% and 3.8%. For the baseline approach (“random”) there are no catalytic active **GOCATs**. Note that we have resisted to show standard deviations and other statistics as these distributions are highly non-Gaussian with high skewness and kurtosis. Moreover, histograms are also not given here as many more data points would have been needed to say something definitive about such distributions. Still, the observed tendency is that the simpler distance-based approach significantly outperforms the **ESP** methods. Additionally, the **ESP** optimization fitness is very similar for both the local and global optimization, with differences probably without statistical significance. Remarkably, the simplest method, “**vdW**”, which actually is not really a translation protocol but only the usual mapping onto the **vdW** surface, performs quite well. This is due to the small deviations between the exposed surfaces for this Menshutkin reaction and almost the same absolute translation and rotation in Cartesian space on both **QM** levels. Thus, the performance of “**vdW**” is expected to deteriorate in more complex translations. Nevertheless, this also strengthens

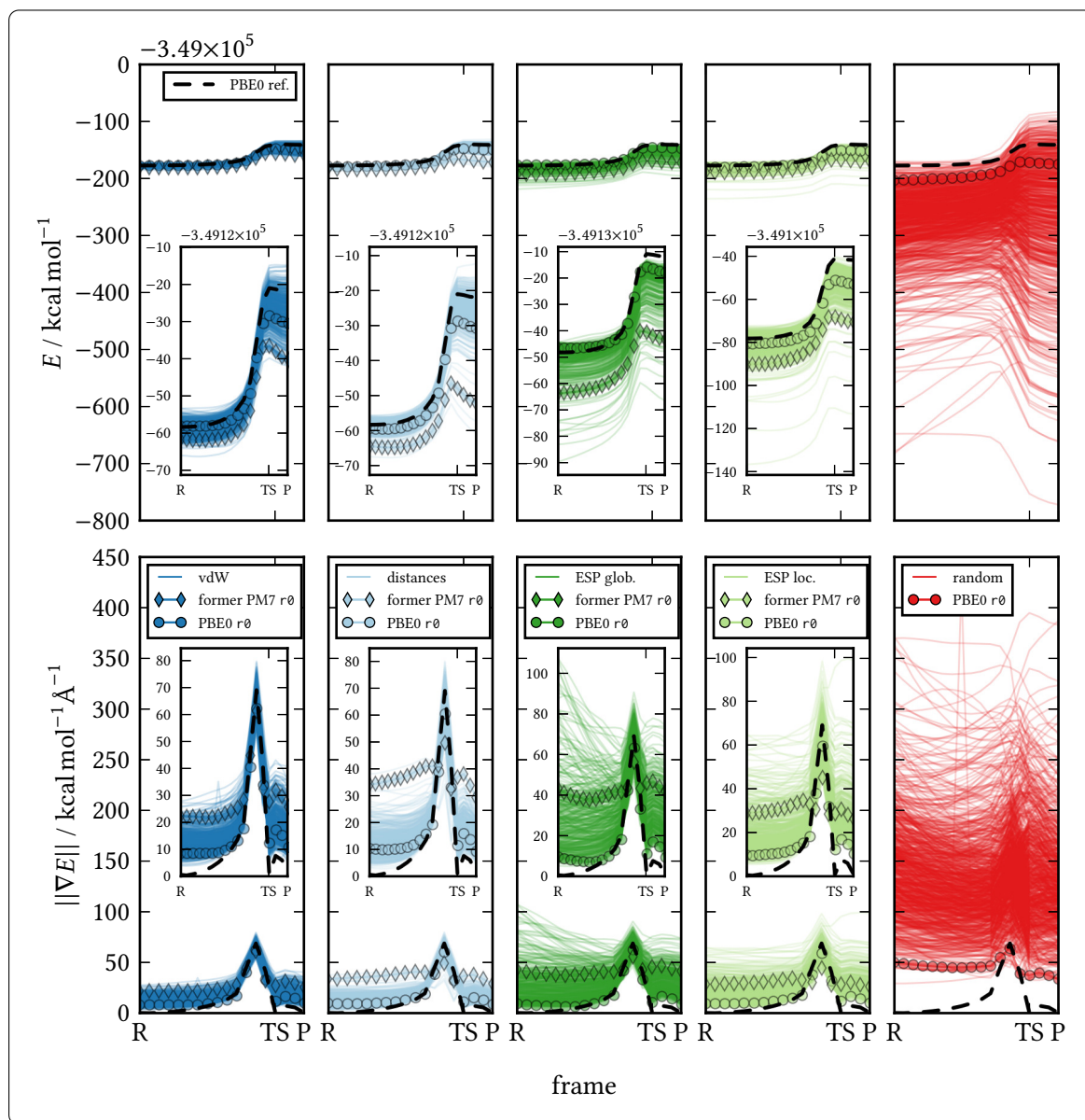


Fig. 6.6: $N_{\text{Ch}} = 10$ (neutral): Corresponding reaction energy profiles and energy gradient norms for all data in Fig. 6.5. “PBE0 r0” is the translation-protocol-specific best GOCAT with the lowest fitness after the translation. These best GOCATs are all pictured below in Figs. 6.8 and 6.9 on p. 169 and on p. 170; “former PM7 r0” is the reaction profile after translation for the former best (extreme of highly symmetric) PM7 GOCAT (shown in Fig. 6.7 on p. 167).

Table 6.2: Descriptive statistics of the fitness of the four translation methods and the baseline approach (random initialization), see Figs. 6.3 and 6.5 on p. 160 and on p. 162 and compare with the corresponding energies in Figs. 6.4 and 6.6 on p. 161 and on the previous page.

		fitness statistics				
	method	N_{GOCAT}	min./max.	median	mean	cat. ratio ^a
$N_{\text{Ch}} = 3$	vdW	378	329.0/30407.3	393.5	687.1	63.8%
	distances	378	331.8/1608.0	386.4	439.0	80.2%
	ESP glob.	377	328.7/4518.6	376.4	884.1	56.2%
	ESP loc.	378	329.7/57658.0	371.9	1375.3	52.9%
	random	461 ^b	444.7/302373.3	36970.6	45884.1	0.0%
$N_{\text{Ch}} = 10$	vdW	474	338.8/8100.1	818.6	1462.6	18.6%
	distances	474	343.6/5546.9	511.1	926.7	29.7%
	ESP glob.	363	340.4/28563.1	2893.1	4302.5	2.8%
	ESP loc.	474	333.4/32765.2	1706.6	2756.2	3.8%
	random	849 ^b	5941.6/NaN	86553.8	108425.9	0.0%

^a Catalytic ratio of GOCATs with fitness values $f < f_{\text{thresh}} = 409.1$; this boundary is the fitness of the pristine path on PBE0/def2-TZVP level of theory.

^b 1000 initializations were calculated. The missing numbers of individuals had at least one frame that could not be converged and consequently those GOCATs were erased. Hence, the open maximum interval is shown (NaN). The other statistics, though, were calculated based on these N_{GOCAT} that could be successfully computed.

the fact that Cartesian information is more important for this translation than combined Cartesian and charge information, at least when no gradients are also included (*vide infra*).

As we were not able to completely trace back the real origin of this, *i.e.*, of the correlations and the causation between ESP and the fitness values, including all sub-terms such as the mere barrier and (more importantly) the gradients, we will give some possible explanations. As can be seen from the energies and gradients of different GOCAT sizes in Figs. 6.4 and 6.6 with $N_{\text{Ch}} \in \{3, 10\}$, the energies themselves for the ESP-based approach are even better than the distance-based approach, but the fitness values are not.

In Table 6.3 on the next page, the averaged gradient norms and energetic barriers are summarized. Here, with ratios around maximal 100%, the ESP translations (locally and globally) find paths with already decreased barriers, ΔE^\ddagger . The translation methods “distances” and “vdW” point around 80–90%. However, the averaged gradient norm of “ESP loc.” and “ESP glob.” are similar and around the threshold of $10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ for $N_{\text{Ch}} = 3$, and suddenly jump to values around $20\text{--}30 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, while the “distances” and “vdW” method leads to smaller values around $10\text{--}15 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$. This restates the fact that “ESP loc./glob.” optimizes proper reaction paths but introduces higher fitness values due to larger gradients at the reaction frames.

Thus, the additional fitness increments will come from the gradient norm penalties. One general observation in the complete electrostatic GOCAT theme is that there is a plethora of Cartesian placements of charges that can lead to (almost) the same ESP at specific points currently observed, *i.e.*, at the $3 \cdot 9$ selected atoms.¹⁵ That is, there is a huge number of

¹⁵The same is true and already well-known in the geometrically “inverse” problem of fitting inner charges

Table 6.3: Descriptive statistics of the energy and gradient norm averaged over the three stationary frames (**R**, **TS**, **P**) for the four translation methods and the baseline approach (random initialization). Compare with the corresponding energy plots of Figs. 6.4 and 6.6 on p. 161 and on p. 163. All energies are given in kcal mol⁻¹ and gradient norms in kcal mol⁻¹Å⁻¹.

	method	type	energy/gradient statistics					cat. ratio ^c
			N_{GOCAT}^a	min./max. ^b	median ^b	mean ^b	σ^b	
$N_{\text{Ch}} = 3$	vdW	ΔE^\ddagger	378	15.2/41.1	34.1	33.9	3.3	84.9%
		$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	378	4.4/69.0	9.1	9.7	4.0	–
	distances	ΔE^\ddagger	378	29.3/40.6	34.7	34.6	1.8	93.9%
		$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	378	3.8/15.8	8.6	9.0	2.5	–
	ESP glob.	ΔE^\ddagger	377	28.5/33.1	30.9	30.7	1.0	100.0%
		$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	377	6.4/38.2	10.1	10.2	2.3	–
ESP loc.	ΔE^\ddagger	378	25.4/38.6	30.9	30.8	1.3	99.7%	
	$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	378	6.2/111.1	9.7	11.5	9.3	–	
random	$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	461	5.2/197.1	71.2	75.0	33.0	–	
$N_{\text{Ch}} = 10$	vdW	ΔE^\ddagger	469	21.0/43.8	33.0	32.9	4.1	85.3%
		$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	474	4.7/36.1	14.5	15.2	5.5	–
	distances	ΔE^\ddagger	468	14.2/42.3	34.3	33.9	3.9	83.1%
		$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	474	3.9/33.5	12.5	13.6	4.9	–
	ESP glob.	ΔE^\ddagger	363	22.1/37.8	30.2	29.9	2.5	99.7%
		$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	363	8.8/77.9	28.4	30.6	13.5	–
ESP loc.	ΔE^\ddagger	474	21.0/35.3	29.7	29.3	2.4	100.0%	
	$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	474	8.5/83.8	20.8	24.0	11.0	–	
random	$\ \bar{\nabla}E_{\text{R,TS,P}}\ $	849	40.7/367.8	129.6	136.3	47.7	–	

^a Number of observations included, varying as, e.g., there might be some very extreme paths without a meaningful barrier to be calculated (see Figs. 6.4 and 6.6).

^b These are the minimum/maximum, median, mean and standard deviation (σ) of the occurring energy barriers (ΔE^\ddagger) or averaged gradient norms ($\|\bar{\nabla}E_{\text{R,TS,P}}\|$) of the N_{GOCAT} evaluated solutions.

^c Catalytic ratio of GOCATs with energy barriers $\Delta E^\ddagger < \Delta E_{\text{ref}}^\ddagger = 37.2$ kcal mol⁻¹; this is the barrier of the pristine path on PBE0/def2-TZVP level of theory.

GOCATs resembling the almost same ESP that will have similar fitness. But this simple correlation is not the whole truth, as gradient norms, for instance, can highly vary when already having the same ESP. This is due to the fact that the gradients themselves include Cartesian information (partial derivative with respect to the $\{x, y, z\}$ -coordinates of the charges), while the ESP itself as a scalar field does not; it only includes the radius or distance. Therefore, in order to improve such descriptions, one maybe would need to either increase the number of points instead of, e.g., the $3 \cdot 9$ of 3 frames and 9 atoms each, and/or one could already include the EF itself at the latter selected points in order to not inadvertently change the gradient norms. Indeed, we often observed that an ESP-optimized GOCAT, although started from a symmetrical individual, reached an unsymmetrical one after ESP

measured from the outside as needed in, e.g., ESP-based atomic charges. In this case, the system is actually underdetermined such that there are multiple realizations of inner charges that lead to a similar or equal accumulated ESP. The same is true for GOCAT's ESP of course and most probably also for the catalytic effect. See the discussion in Section 2.1.1.2

optimization with rather different charge values, $\{q_i\}$. Then, opposing charges will not compensate for the gradients but might introduce the same ESP and yet a different EF at the respective atom.

As a consequence, as there are numerous ways of Cartesian representations, the simple distance-based approach relaxing to the nearest Cartesian local optimum without changing the charge values might be more effective than the ESP method that actually only uses the ESP as a “template”, but might find completely different Cartesian GOCATs with some loss of information due to the rough grid-based view without incorporating the EF.

With regard to the depictions in Section 2.1.1.2, also the density of calculated ESP grid points is usually notably higher for PD charges, with roughly up to a few hundred points per atom^[141] or one to many points per Å² at multiple shells.^[154,453]¹⁶ Moreover, different protocols, e.g., similar to RESP,^[153,154] could be used as “Occam’s razor” incorporated for simplifying the solution space.

The highly symmetrical PM7 best rank, $r\theta$, is shown again in Figs. 6.7(a) and 6.7(b) on the following page; this individual was discussed around the publication’s Fig. 8 on p. 138. All three protocols (“distance”, “ESP glob.” and “ESP loc.”) deliver rather unsymmetrical GOCATs after translation (Figs. 6.7(c)–(h)). Moreover, the ESP range of the $r\theta$ GOCAT is quite high with a difference of $\Delta\varphi_{\text{ESP}} \approx 40 \text{ kcal mol}^{-1} \text{ e}^{-1}$ between the Cl and C atoms. This rather extreme case was shown to be an outlier but a very good one on PM7 (see the publication on pp. 137f.), and, *eo ipso*, it can be assumed that this individual is harder to be translated onto PBE0. A reason could be that such a rather extreme embedding simply is not performing well on a higher level of theory, with, e.g., also a more extensive basis set and a better QM/MM coupling, or that the translation error produced by the methods perishes the candidate solution’s quality; every small translation error that is introduced, such as not compensating gradients due to asymmetrically opposing charges, will be amplified when having to treat an extreme GOCAT. By contrast, the well-performing GOCATs on PBE0 all stem from ranks in the region of 3000–6000. These mediocre GOCATs on PM7 that happen to be the best ones on PBE0 after the translation are all pictured in Figs. 6.8 and 6.9 on p. 169 and on p. 170. They all have a smaller $\Delta\varphi_{\text{ESP}} \approx 20$ and thus seem to be less an issue for the translation protocols.

As a conclusion of this Section, the following three points can be summarized: (1) *Just* translating the PM7 GOCATs without any further GOCAT optimization with regard to the catalytic effect *already* shows meaningful catalytic effects on PBE0 level of theory for many individuals, *i.e.*, the results based on the additional SQC approximations (Sections 2.2.3 and 2.4.3) can be transferred. (2) The ESP optimization should be extended to include more points and/or the EF. Alternatively, one could include some simplification pressure, as mentioned in Section 2.1.1.2, or enforce symmetry right away. (3) Local optimization of ESP could be sufficient, and the global setting is of greater computational expense and not needed for translation.

¹⁶Note that in other GOCAT compression applications—not shown here—such ideas were already pursued and the emerged positioning seemed to be very dependent on subtle differences of the target ESP and on the number of points included. Which impact this has on the translation between levels of theory, has to be investigated further.

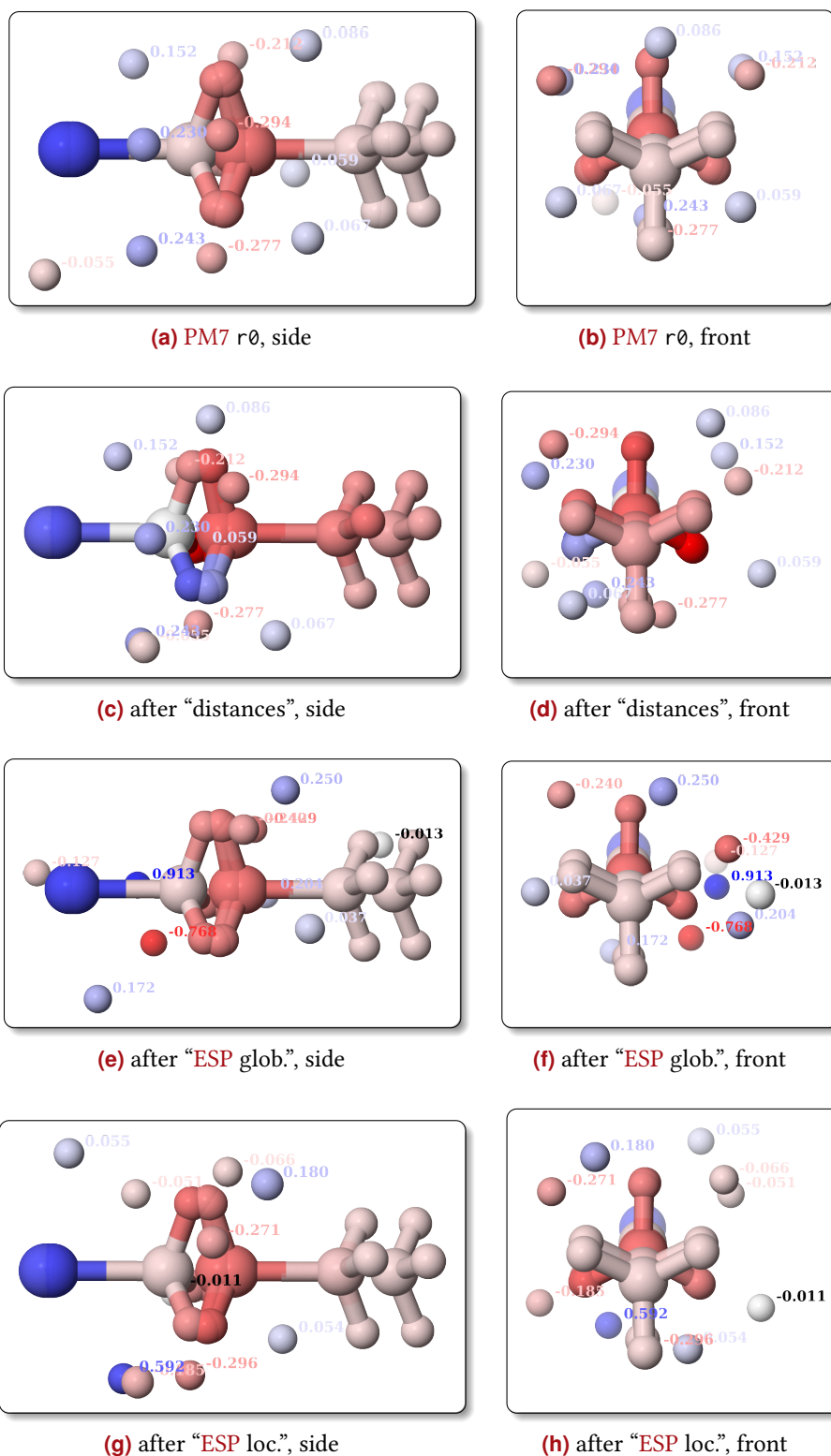


Fig. 6.7: $N_{\text{Ch}} = 10$ (neutral): Exemplary GOCATs as starting individual on PM7 (highly symmetric best rank r_0), Figs. (a)–(b), and translated by the three methods: “distances”, “ESP glob(ally)” and “ESP loc(cally)”. Charges and atoms are colored from red to blue for $q_i \in [-1.0, +1.0]$ e and $\varphi_{\text{ESP}} \in [-38.8, +38.8]$ kcal mol⁻¹e⁻¹.

Note that the results of this Section show fully non-optimized **GOCATs** with respect to the catalytic effect. The **GOCATs** for both “distances” and “vdW” together for the $N_{\text{Ch}} = 10$ size constituted the very same starting population of the publication for the **GA** optimization on the **DFT** level, discussed on p. 141.

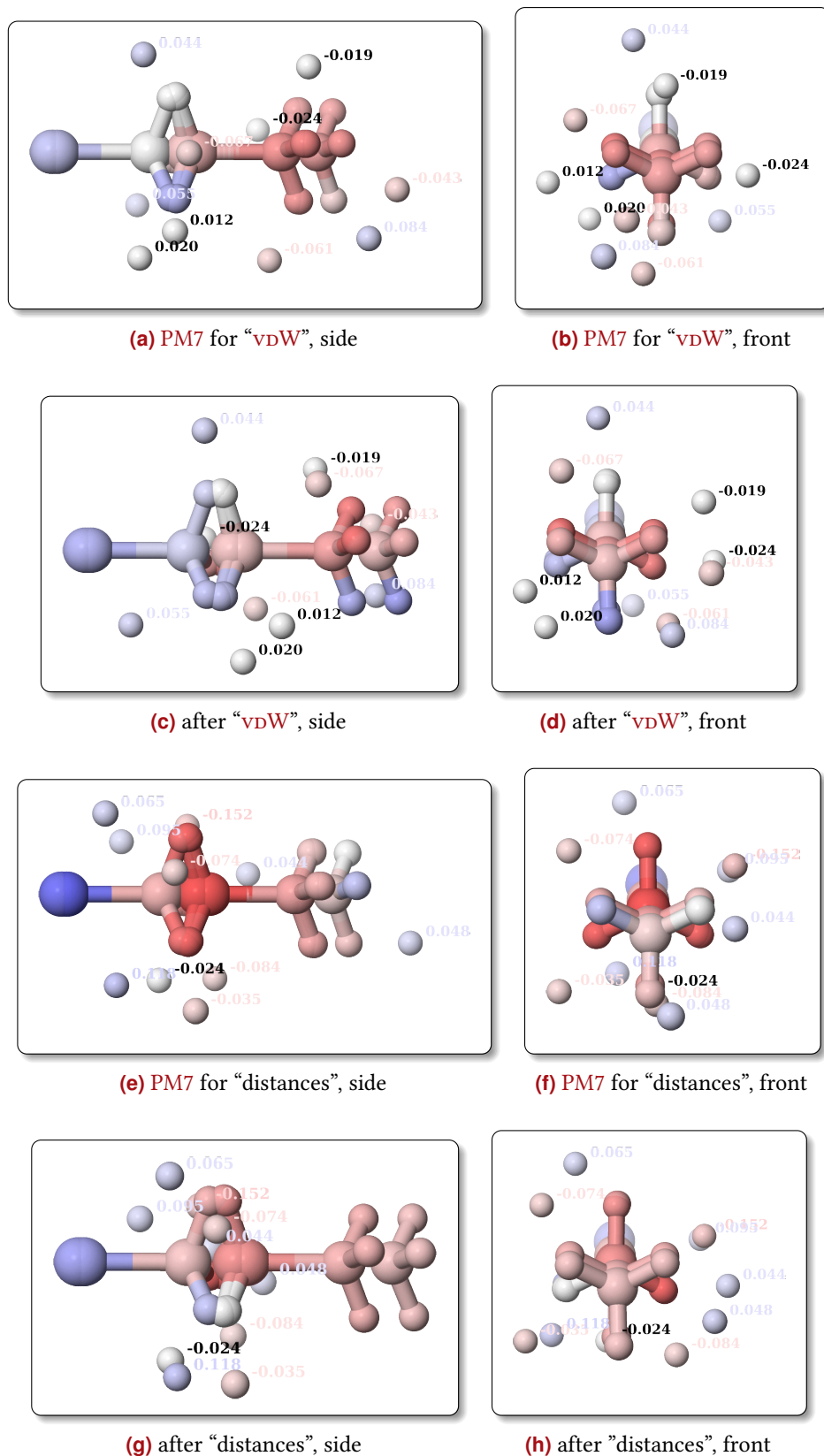
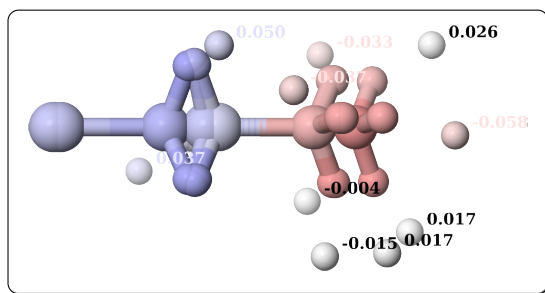
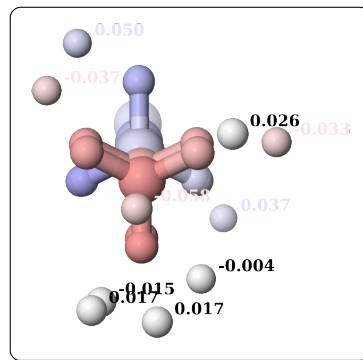


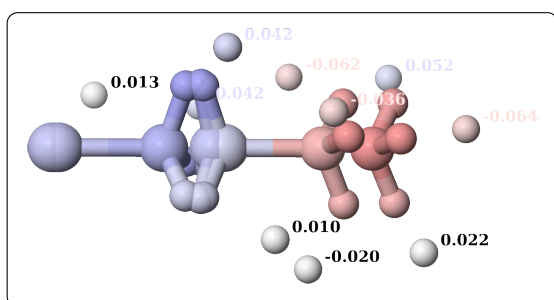
Fig. 6.8: $N_{\text{Ch}} = 10$ (neutral): GOCATs as starting individual on PM7, Figs. (a)–(b) and Figs. (e)–(f), and their translation to PBE0 based on the method illustrated. These are the case-specific GOCATs that happen to give the lowest fitness on PBE0 (dotted reaction paths in Fig. 6.6 named “PBE0 $r\theta$ ”). Charges and atoms are colored from red to blue for $q_i \in [-1.0, +1.0]$ e and $\varphi_{\text{ESP}} \in [-13.6, +13.6]$ kcal mol⁻¹e⁻¹.



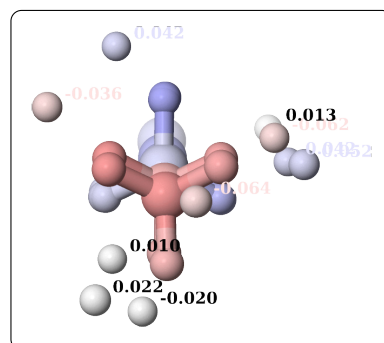
(a) PM7 for “ESP glob.”, side



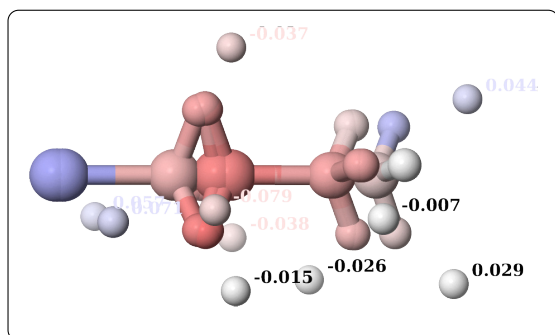
(b) PM7 for “ESP glob.”, front



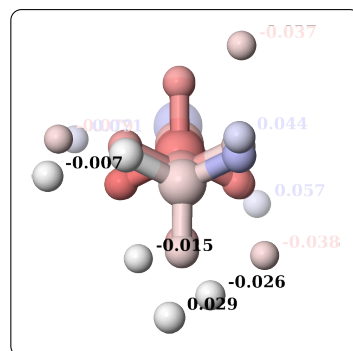
(c) after “ESP glob.”, side



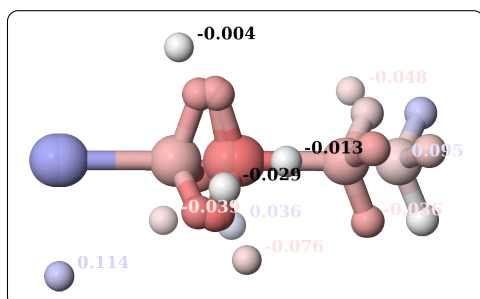
(d) after “ESP glob.”, front



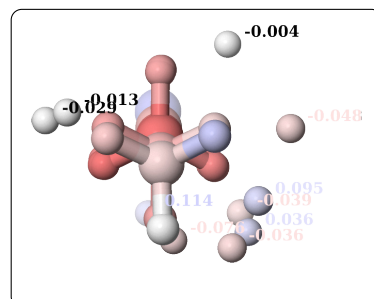
(e) PM7 for “ESP loc.”, side



(f) PM7 for “ESP loc.”, front



(g) after “ESP loc.”, side



(h) after “ESP loc.”, front

Fig. 6.9: $N_{\text{Ch}} = 10$ (neutral): The other two translation methods complementary to Fig. 6.8 on the preceding page (see the caption there).

6.3.3 Correlation Studies

Complementary to the simple catalytic impact of any polar surrounding on the Menshutkin reaction discussed in the paper, this Section sheds light on this matter again from the perspective of recently emerging pure electrostatic catalysis, which is also detailed later in Section 7.1; the simple trend for the Menshutkin reaction can serve as a basis for comparison there.

6.3.3.1 Primer on Electrostatic Catalysis

General electrostatic catalysis can be estimated by the equation for the barrier lowering as^[122–124,126,127]

$$\Delta\Delta E^\ddagger = - \left[(\mathbf{F}_{\text{env,TS}} \cdot \boldsymbol{\mu}_{\text{TS}}) - (\mathbf{F}_{\text{env,R}} \cdot \boldsymbol{\mu}_{\text{R}}) \right], \quad (6.2)$$

where $\mathbf{F}_{\text{env},\{\text{TS},\text{R}\}} = -\nabla\varphi_{\text{ESP},\{\text{TS},\text{R}\}}$ is the EF at the corresponding frame, either **R** or **TS**, and $\boldsymbol{\mu}_{\{\text{TS},\text{R}\}}$ its (bond) dipole moment. Assuming overall neutrality of the molecule, the energy difference within an electrostatic environment can be approximated by perturbation theory to the first order as an interaction of the (homogeneous) electric field and the corresponding (molecular) dipole moment, $\Delta E = -\mathbf{F} \cdot \boldsymbol{\mu}$ (cf. Section 2.3.2). Especially, vibrational Stark shifts within an electric field were estimated in this way.^[454–456] In this Thesis, we deliberately describe this energy influence as a *potential* energy change excluding entropy effects,^[127] although, in the given references, it is usually used as a direct impact on the free energy barrier, $\Delta\Delta G^\ddagger$. In some reactions, the main electron density re-configuration can sufficiently be approximated by a *local* electron rearrangement during the reaction such that the local electric field at specific functional groups and their local bond dipoles can be used, instead of the full molecular dipole moment.^[121,122,457] Besides, it is assumed that a (small) uniform electric field and this dipole approximation is adequate. Indeed, a projection of the dipole moment and the EF onto the main reacting bond(s) usually suffices.^[123,124,126,127]

In this simple picture, electrostatic catalysis can be bounded by two limiting cases, assuming one common electric field of an environment, \mathbf{F}_{env} , for all structures emerging during the chemical reaction—which is completely in line with the **GOCAT** model. These are illustrated in Fig. 6.10 on the next page:^[122]

- First case (shown in Fig. 6.10(a)): Here, the dipole moment of the molecule does not change the direction but its magnitude during the reaction—or more precisely its Euclidean norm. Then, each stabilization of **R** will *automatically* also stabilize the **TS** structure. If the **TS** is stabilized stronger due to $\|\boldsymbol{\mu}_{\text{R}}\| < \|\boldsymbol{\mu}_{\text{TS}}\|$, we have a catalytic effect, *i.e.*, a barrier decrease of Eq. (6.2), which, in this case, simplifies to $\Delta\Delta E^\ddagger = -\|\mathbf{F}_{\text{env}}\| (\|\boldsymbol{\mu}_{\text{TS}}\| - \|\boldsymbol{\mu}_{\text{R}}\|)$. Also, a reaction field as the electric field by an arbitrary (polar) solvent—assuming a perfectly equilibrated solvent phase at 0 K—onto the solute will thus lead to a stabilization and, accordingly, to the solvation energy, by a field in the same direction as the solute’s dipole moment. Excluding thermal effects, the solvent molecules will be aligned in a way to minimize the total energy including the solvent–solute interactions of, *e.g.*, the stable starting structure

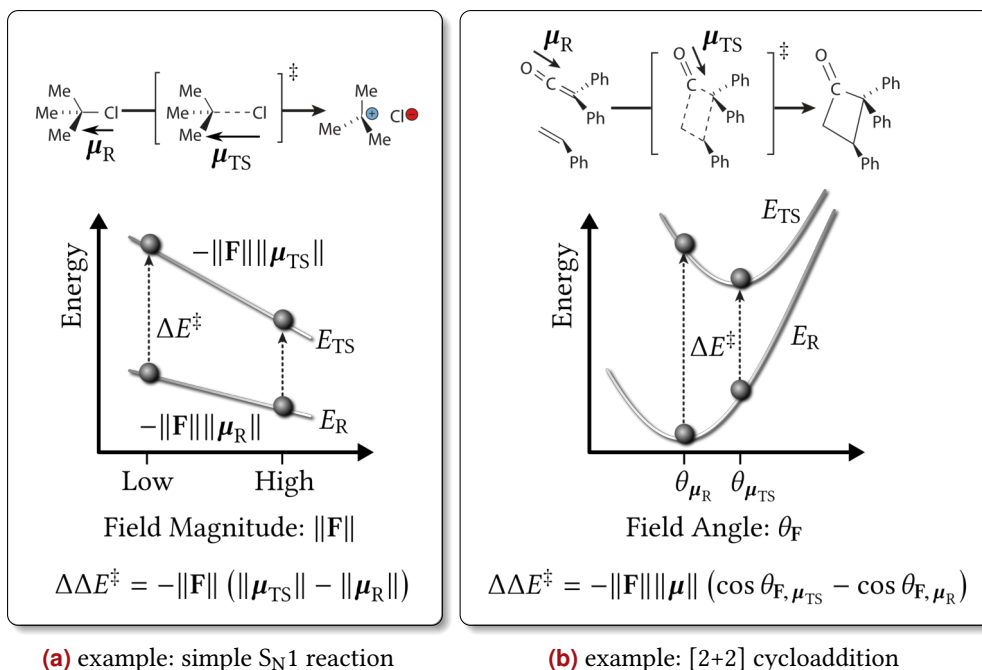


Fig. 6.10: Corner cases of catalysis by electric fields of environments. In Fig. (a), the first heterolysis step of an S_N1 reaction to the carbocation greatly increases the dipole moment from **R** to **TS** (and **P**), but the direction stays the same. In Fig. (b), by contrast, the (local bond) dipole (of the C=O group) re-orient, but the magnitude stays (approximately) the same, illustrated by a ketene electrocyclic reaction. The sign convention here is the common one for the electric dipole moments to point from minus to plus, electric fields from plus to minus. Perfectly aligned fields will thus be parallel to the dipole moment. Stabilizing interactions will then be negative for ΔE due to the negative prefactor in Eq. (6.2). The pictures are adapted from Ref. [122].

R. By this, automatically the **TS** is stabilized more, *i.e.*, the solvent does not have to re-orient/re-equilibrate, and the barrier decreases.

- Second case (Fig. 6.10(b)): If the dipole moment does not change its magnitude, but its direction upon forming the **TS**, the environment has to be preorganized¹⁷ such that it is already adjusted to the **TS** situation, *i.e.*, mirroring its (complementary) electrostatics. Each solvent that is stabilizing the **R** structure will fail to stabilize the **TS** more in order to reach a barrier decrease, assuming again a “frozen” solvent perfectly attached to **R**. Hence, an environment (other than a solvent) needs to “anticipate” the electrostatics of the **TS** in order to provide an electric field for optimal **TS** stabilization and thus a barrier decrease. Here, Eq. (6.2) becomes $\Delta\Delta E^\ddagger = -\|\mathbf{F}_{\text{env}}\| \|\boldsymbol{\mu}\| (\cos \theta_{\mathbf{F}, \boldsymbol{\mu}_{\text{TS}}} - \cos \theta_{\mathbf{F}, \boldsymbol{\mu}_{\text{R}}})$.

Certainly, real reactions can often not be subsumed as easily under one of these limiting cases as both the dipoles’ orientation and magnitude will change at the same time. Moreover, the possibility of inhomogeneous electric fields of an environment will complicate matters and lead to the breakdown of a simple point dipole description, when multiple (also distant) charge rearrangements happen during the reaction synchronously and further

¹⁷Note that this is what WARSHEL^[112] dubbed as “preorganization” of an environment mentioned in the Introduction in Chapter 1, more specifically of enzymes, which was re-emphasized multiple times since then.^[114–117,458]

“directionality” of the electrostatics is needed, *i.e.*, higher order moments and mixed higher-order effects should be included. Lastly, there is usually a correlated motion between the pure internal reaction and the outer environment and not a purely fixed or static single EF. Despite these apparent deficiencies of this simple description, it was often used in practice by different theoreticians and also in experimental settings and indeed showed linear correlations with the otherwise computed or measured barriers.^[122–124,126,127,455] This procedure may thus serve as baseline for the explanation or cause of the final barriers within electrostatic surroundings.

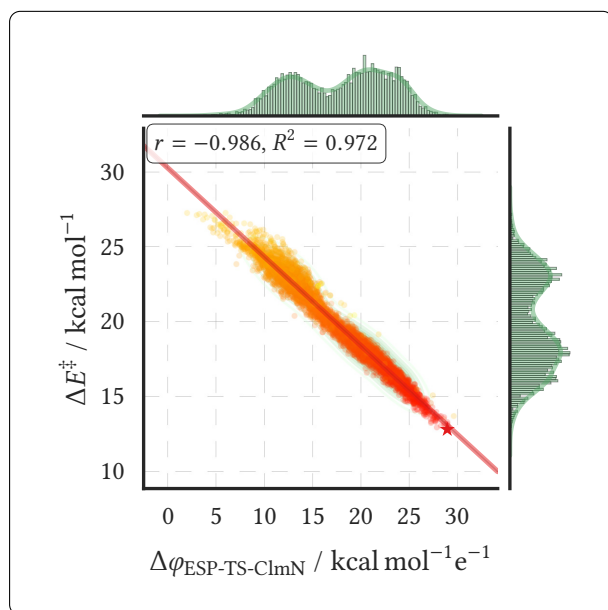
6.3.3.2 Menshutkin Reaction as Limiting Case

The Menshutkin reaction as thoroughly discussed in the publication of this Chapter (Section 6.2) had the well-known clear tendency of being catalyzed and also becoming exothermic by increasing polarity of the solvent, which was the actual reason for selecting this reaction for the proof-of-principle study in the first place. With regard to the corner cases above, this neutral S_N2 reaction clearly can be assigned to the case of Fig. 6.10(a). Due to the high symmetry during the whole reaction, belonging to the point group C_{3v}, and, accordingly, the linear reaction coordinate, mainly the dipole magnitude but not its direction changes when the products are formed. Remember that the P structure is charge-separated but neutral overall (Cl[−] ⋯ H₃C–NH₃⁺) and that the dipole essentially is the displacement of the charged particles (“fuzzy” electrons, and fixed nuclei in BO-QM) from an origin, which is arbitrary in the case of a neutral system. Thus, the P and already the TS structures have heavily increased dipole moment magnitudes (in Debye): $\|\mu_R\| = 4.7 \text{ D} < \|\mu_{TS}\| = 14.2 \text{ D} < \|\mu_P\| = 17.0 \text{ D}$.

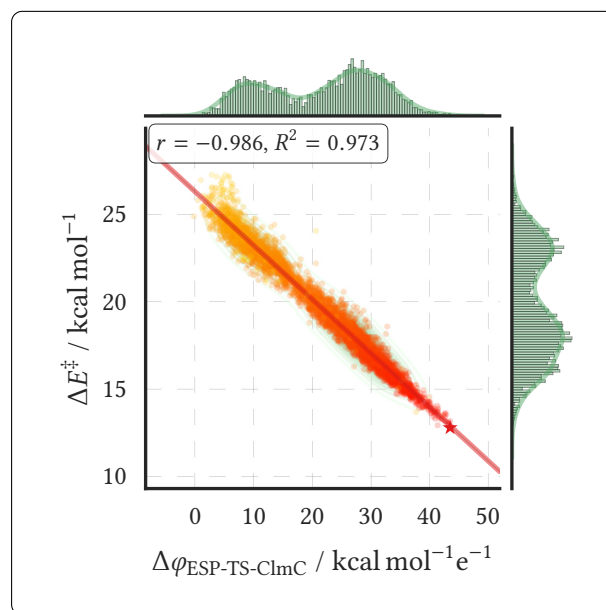
In this regard, the total population of $N_{\text{Ch}} = 10$ (neutral, vdW) discussed in the publication is investigated again. This leads to the correlation plots of several properties in Fig. 6.11 on the following page. $\Delta\varphi_{\text{ESP}}$ denotes the differences of ESPs at the indexed atoms of the TS frame (compare with Figs. 6.7(a) and 6.7(b) on p. 167): *E.g.*, “ClmN” labels the difference of $\varphi_{\text{ESP,Cl}}$ minus $\varphi_{\text{ESP,N}}$, which maps to the overall ESP influence onto the (prebuild) anion and cation. Also, more intermediate cases are shown such as “ClmC” which terms the ESP difference between the (left) Cl and the (middle) C atom and “CmN” for the one between the (middle) C atom and (right) N atom.

Apparently, the barrier is highly correlated with any of the properties. With Pearson correlation coefficients of $r \geq 0.931$ or the coefficients of determinations of the simple linear regression of $R^2 \geq 0.866$, a high percentage of the variance in the dependent variable, the final barrier ΔE^\ddagger , can be “explained” by the difference in the ESP values at the respective atoms.¹⁸ By simply increasing the total ESP difference or increasing the total electric field from Cl to N (Fig. 6.11(a)) or between Cl and C (Fig. 6.11(b)) the barrier is decreased in the range of $\Delta E^\ddagger \in [12.77, 27.24] \text{ kcal mol}^{-1}$ while $\Delta\varphi_{\text{ESP-TS-ClmN}} \in [2.02, 29.62] \text{ kcal mol}^{-1} \text{ e}^{-1}$ or $\Delta\varphi_{\text{ESP-TS-ClmC}} \in [-0.13, 43.68] \text{ kcal mol}^{-1} \text{ e}^{-1}$. As already seen from the last number, the

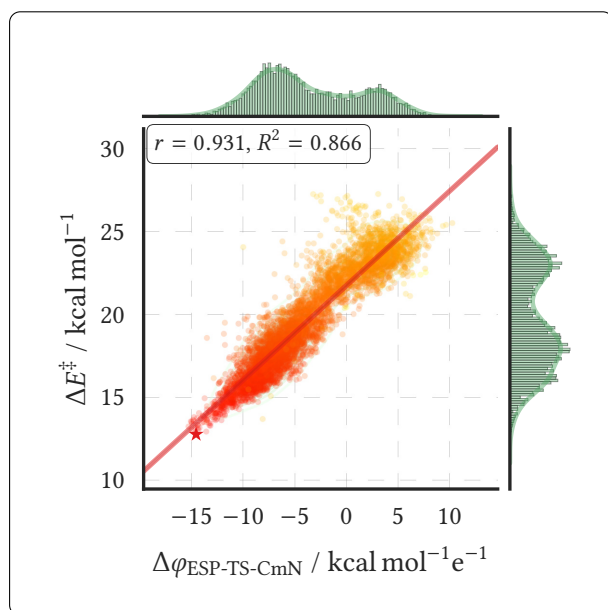
¹⁸Note that in simple linear regression with an intercept, $r^2 = R^2$ hold (not to be confused by the two different symbols). The former merely adds the information of the sign for positive or negative linear correlation. Despite this, both values were always plotted and analyzed in multiple thousand times for other reactions/settings.



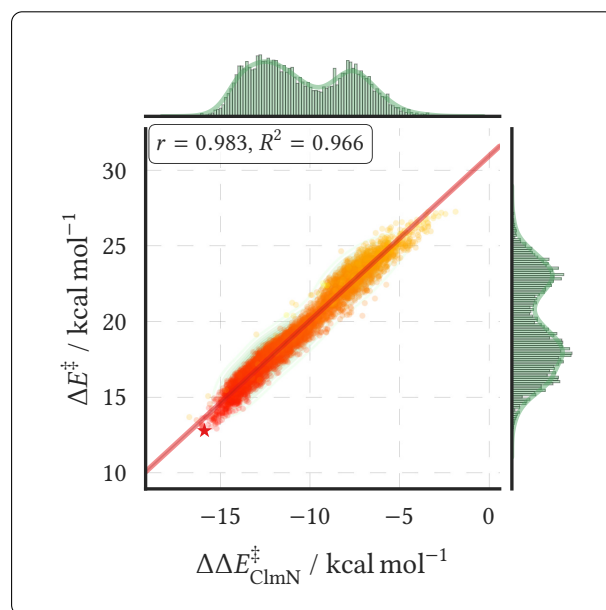
(a) barrier decrease with “total” $\Delta\varphi_{\text{ESP}}$



(b) barrier decrease with first half of $\Delta\varphi_{\text{ESP}}$



(c) barrier decrease with other half of $\Delta\varphi_{\text{ESP}}$



(d) estimated barrier decrease from Eq. (6.2)

Fig. 6.11: $N_{\text{Ch}} = 10$ (neutral), PM7 for the Menshutkin reaction of all candidate solutions ($N_{\text{GOCAT}} = 5388$): The barrier, ΔE^{\ddagger} , is plotted against the difference of ESP (the voltage) at the TS frame of different atom-pairs, Cl minus N, Cl minus C and C minus N. The colormap of the scatter points is changing from yellow to red for high to low fitness values, and the best rank, r_0 , is marked with a red star (cf. Figs. 6.7(a)–(b) on p. 167). Additionally, a simple linear regression is plotted with a red line and corresponding histograms of the data points are given in the margins. $\Delta\Delta E_{\text{ClmN}}^{\ddagger}$ is the *estimated* barrier decrease by Eq. (6.2).

C atom is negatively embedded at the contact plane of the ions leading to the positive correlation, *i.e.*, less negative ESP at N, in Fig. 6.11(c) that even lies in both negative and positive domains: $\Delta\phi_{\text{ESP-TS-CmN}} \in [-15.26, 10.27] \text{ kcal mol}^{-1} \text{ e}^{-1}$.¹⁹ By adding the “third” dimension, the colormap, the correlation is stressed even more. The “star” denoting the $r\theta$ (Figs. 6.7(a) and 6.7(b)) is one of the most extreme cases and found with the lowest barrier and highest ESP difference. Finally, the resulting barrier, ΔE^\ddagger , can simply be traced back to the projected electric field and molecular dipole moment onto the direction vector between Cl and N (all { Cl, C, N } are on one axis): $\Delta\Delta E_{\text{ClmN}}^\ddagger \in [-16.71, -1.86] \text{ kcal mol}^{-1}$ is the calculated barrier decrease due to Eq. (6.2) using the polarized molecular dipole moments for the projection.²⁰ Thus, Eq. (6.2) can describe about $R^2 = 0.966$ of the total barrier variance in the full Menshutkin pool leading to $\Delta E^\ddagger - \Delta E_{\text{ref}}^\ddagger = \Delta\Delta E^\ddagger \in [-17.68, -3.21] \text{ kcal mol}^{-1}$. The latter numbers describe the resulting real barrier decreases of the GOCATs, $\Delta\Delta E^\ddagger$, and are even smaller $-17.68 < -16.71$ than the estimated ones. Deviations can always stem from the inhomogeneities of the fields as asymmetry along the axis at the H atoms excluded in such a simple regression and create the “noise” seen in the Figures. Compare also with the electric fields given in the appendix for the plain COSMO case and $r\theta$ with field components along the Cl–C–N axis, but more local inhomogeneities (Figs. A.5 and A.6 on pp. 270f.).

In summary, these correlations again underline the very simple nature of the impact of electric fields on the Menshutkin reaction, and the expected trend is very well realized by (almost) all the GOCATs found.²¹ However, such simple trends for all solutions found after the GA is rather an exception. Similar discussions will also follow in Chapter 7 for the case of a DA reaction, a reaction with both a changing dipole moment direction and magnitude.²²

¹⁹ These values and corresponding histograms are discussed on p. 137 in the publication.

²⁰ $\Delta E = -\mu F$ with polarized μ within the GOCAT (QM/MM) is supposed to include all higher-order moments in the picture of perturbation theory, but the EF is highly averaged, *i.e.*, using one average field between Cl and N regardless of the local changes in magnitude or direction of the field vectors.

²¹ One follow-up investigation of the current author was the simple chloride exchange $\text{Cl}^- + \text{CH}_3\text{Cl} \longrightarrow \text{ClH}_3\text{C} + \text{Cl}^-$ then. In this case, the R and P structures are formally equal except for the fact of differing absolute coordinates (left, right) that are important when having an anchor point of a GOCAT. In this case, the molecular dipole moment magnitude of the TS structure ($[\text{Cl}\cdots\text{CH}_3\cdots\text{Cl}]^-$) must be smaller than the ones of the R or P structures. Notwithstanding, a GOCAT can be designed with almost just an *orthogonal field* to the “reaction axis” (*cf.* Section 7.1 where this term is introduced). The GOCAT shows ESP differences at the H atoms in the middle, resulting in equal energies and an equal ESP at R and P. This can thus be thought of as a full “anticipation” of the TS electrostatics (not shown).

²² Actually, multiple (ten) full sets of optimizations for the ketene reaction of Fig. 6.10(b) were calculated, but these still need to be thoroughly analyzed/understood. As the GOCAT model uses fully aligned frames of the reaction path, simple descriptions of “molecular dipole moment directions” can be misleading and even varying between levels of theories. In fact as a result, the ketene reactions did not change any overall dipole moment direction so much on PM7. Therefore, the DA project including the adaptive fitness function (Algorithm 3.2) was tackled.

Adaptive Globally Optimal Catalysts

Manifest criticism that could be directed at the **GOCAT** model introduced and utilized so far (*cf.* Section 6.2) could be its severe approximate nature that seems quite far from what actual *concrete* catalysts (can) do. Having to start *somewhere*, the general ingredients were motivated already in the Introduction (Chapter 1), but the question arises whether some of the model restrictions can be relaxed in order to reach the next step on the way of the endeavor of “fully automatic catalysis design”, which after all is the ultimate goal—despite the challenge of being quite ambitious. Therefore, apparent deficiencies present in the **GOCAT** model were discussed already in Section 6.2 (p. 142). Accordingly, this Section tries to investigate a route of relaxing the restrictions bearing on fixed or preoptimized reaction paths (**MEPs**) as input, treating single-step reactions with one clear **TS** only and, by the same means, bearing on a fixed reaction mechanism.¹

Fortunately, there is emerging new research focusing on electrostatic catalysis *per se*, by an anyhow externally applied **EF**. This reasserts also the importance of pure, “abstract” **EF**-based catalysis. In this context, also the fully electrostatic **GOCATs** that are still optimized during *this* Thesis might gain ground by providing insights and realizations of globally optimal electrostatic embeddings.

Next, the recent research is briefly recapitulated and, subsequently, further studies regarding mechanistic changes during **EF** impact follow for a **Diels–Alder (DA)** reaction.

7.1 Overview on Recent Electrostatic Catalysis

Recently, several efforts of understanding and modeling of catalytic effects evolved into a discipline that could be termed *electrostatic catalysis* and was reviewed in both the theoretical^[268,459] and the experimental domains.^[122,460] In this regard, so-called **oriented external electric fields (OEEFs)**^[459] play a vital role as, generally, dipolar uniform fields (or

¹ Still, there are, naturally, other important improvements needed in further future adaptations (see Section 7.5 and especially the prospects in Section 8.3).

the ones produced, e.g., at scanning tunneling microscope (STM) tips or charged surfaces) along a fixed given direction affecting the molecular system.^[268] In the cited reviews, it was demonstrated for different reactions (e.g., Refs. [120, 461–466]) that an OEEF along the direction of electron reorganization catalyzes *nonpolar*² and *non-redox* reactions by orders of magnitude. This direction was called the *reaction axis*, as the direction in which the “electrons move for building bonds” from R to P. Fields in other directions can then control regio- and stereoselectivity. By increase of the field strength, this impact can induce a mechanistic turnover with emerging ionic species. Most of these influences can be understood by a field-induced stabilization of ionic structures.³ Recently, there were also time-dependent investigations (*ab initio* MD under periodic boundary conditions, including metadynamics) of the synthesis of methane and formaldehyde or glycoaldehydes under EF-influence, though, under otherwise ambient conditions, with observed changes of the reaction network.^[469–471]

This recent research interest also extends to enzymatic catalysis,^[120,123–128] which leads back to the origins of electrostatic proposals of how enzymes work of WARSHEL in, e.g., Refs. [112–115]; in this context, also the term **local electric field (LEF)** was coined^[268] classifying the unique surrounding an enzyme can (locally) generate. Yet, in this context, the vivid debate should be emphasized again, still continuing until today, about the origin of enzymatic catalysis, which can be divided into two major groups: On the one hand, the proponents of *electrostatic* catalysis in a preorganized polar enzymatic environment stabilizing the TS more than R, and on the other hand the ones of general *dynamical effects*, leaving the area of thermally driven (*i.e.*, reaching non-TST-based) effects and arguing that dynamics play a pivotal role. WARSHEL most vigorously represents the first (electrostatic) view, as argued in detail in different Refs. [115–117] against the other view (see, e.g., Refs. [69, 118, 119], the introduction of Refs. [122, 127] and references therein).

As pointed out already in the Introduction of this Thesis (Chapter 1), we do not want to take sides in this debate: If an electrostatic preorganization sufficed to describe the full power of enzymes, this would doubtlessly underline the importance of introducing and optimizing (abstract) point charge GOCATs. If not, this would stress the fact that naturally occurring catalysts need more specific dynamical effects and so should a further extended GOCAT model. However, note that the very recent research regarding non-enzymatic electrostatic catalysis of OEEF puts forward *abstract* embeddings for pure electrostatic effects, which is in line with the current GOCAT model.

Moreover, FRIED and BOXER have developed vibrational Stark shift spectroscopy in order to *measure* naturally occurring electric fields in enzymes.^[121,122,457] From these works, the treatment of local dipoles stems, which was already explained in Section 6.3.3.1. They immobilize molecules of interest, can apply a uniform EF onto these and measure the (infrared)

² The system can even be completely non-polar but polarizable during the reaction, *i.e.*, more so at the TS structure.

³ One of the key-players is SHAIK in the theoretical domain, who also explains many of the found chemical changes due to an EF by the use of modern valence bond theory which SHAIK reviews in, e.g., Refs. [467, 468]. Regarding the research in the EF topic, the supporting information of Ref. [268] can be consulted with citing >300 more References.

spectrum. The field shifts the involved (vibrational) states differently due to different dipole moments of the latter. By this, the applied field affects the transition frequencies in a generally linear fashion with respect to the difference in dipoles of the involved states. This can then be used as calibration for an “inverted” experiment, where probe molecules of known frequency shifts, carrying distinct probe functional groups of, for instance, C=O or C≡N, are embedded into the new environments such as enzymes to measure the frequency shifts by the new surrounding. With complementary MD computations for projecting the electric fields onto the local probe group directions and with help of further solvatochromism calibrations, they can map the vibrational frequency shifts to absolute electric fields. For instance, in this way an extreme EF that ranges up to $\sim 150 \text{ MV cm}^{-1}$ (or $34.6 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ in the units used below) was measured in the enzyme ketosteroid isomerase in the active site.^[121]

This methodology seems promising, and also in a theoretical domain such local projections of EF onto the important bonds were applied,^[122–124,126,127,455] which generally is in line with the aforementioned reaction axis rule. That is, the direction of electron density changes during the bond making and forming process is crucial for the assessment of electrostatic catalysis.

Apart from that, again mainly in the experimental domain in combination with theory, COOTE *et al.* have investigated the role of EFs and how to harness them for catalysis *in practice*, reaching even the daily laboratory regime.^[460] In a single-molecule setting, electrostatic catalysis for a DA reaction was experimentally proven to occur using an STM “blinking” experiment.^[472] This has shown *qualitative* agreement with the computational studies in the same Reference for the studied DA systems.⁴ Here, both the electrostatic effect must be differentiated from other effects, such as electron tunneling itself, and more importantly for bimolecular reactions, the molecules must be tethered and the EF direction must be precisely controlled. Likewise this bond-forming proof-of-concept study under EF, but using both an electrostatic and electrochemical impacts, bond breakage was also investigated in another study.^[473]

As STM is undeniably not a “high-throughput” approach for EF catalysis for a real synthesis, besides some other approaches using surfaces of electrodes for so-called **interfacial electric fields (IEFs)**, the main further research of COOTE is focused on charged functional groups for local oriented EF with, *e.g.*, a pH switch to control the charged group. This was also named **designed-local electric field (D-LEF)** in Ref. [268] and requires systems where mainly electrostatic effects of specifically introduced functional groups dominate, and not others as, *e.g.*, conjugative effects of protonations etc. They showed, for instance, H-atom transfer reactivity enhancements due to a D-LEF^[474] and a high catalytic influence onto a DA reaction using such local fields, including also *endo/exo* selectivity,^[475] under practical solution-phase conditions (in less polar ones, of course).

In order not to miss out another broad topic in its own—while the mentioned STM

⁴ With all types of approximations that are needed to model the same STM experiment in theory, for instance, approximations to model gold tips, entropic effects in an STM, isolated barriers in an STM for the reactions (without interactions), the real field (on an atomistically rough surface) and solvent effects.^[472]

experiments already point into that direction—there is also much research going in the field of **EFs** for *heterogeneous* catalysis.^[476]

For the remainder of this Section, the mentioned mechanistic changes due to an **EF** are most essential as well as having another probe, the dipole moments and the **EFs** projected onto the central electronic flow directions (reaction axis). These can help to quantify the expected electrostatic catalytic effects and, furthermore, support to discern occurring influences of **GOCATs** from the ones that are possible by simple uniform fields, which are evaluated in a point dipole description.

7.2 Diels–Alder Reaction

One early attempt on such **EF**-based catalysis was the study of a **DA** reaction.^[463] Here, the two synchronously formed C–C-bonds define the reaction axis. A field in this direction was shown to highly catalyze an archetypal **DA** reaction (in a slightly polar variant). Moreover, fields orthogonal to the reaction axis could be used as handle to control the *exo* vs. *endo* barriers differently, which opens the possibility of better kinetic control of the outcomes of both diastereomers. This basic catalytic effect was, as mentioned above, already validated in an **STM** experiment for an **OEEF**,^[472] and also in a **D-LEF** setting with pH switches on distant (non-conjugated) groups of the involved species.^[475] So far, this **DA** reaction even culminated in another study of full *enantioselective* control of the outcome using **OEEFs** for **DA** reactions,^[465] for systems where this plays a role, contrary to the more simplified one in the following that is equal to the first study (Ref. [463]).

This **DA** reaction is shown in Fig. 7.1. It illustrates both the *exo* and *endo* diastereomers of the [4+2] cycloaddition reaction of cyclopentadiene and maleic anhydride. The anhydride as dienophile with its electron-withdrawing properties accelerates the reaction when compared to the plain barebone reaction without any additional functional groups. This property is also illustrated with the partial charge, $\delta^{\{+,-\}}$, as used in organic chemistry. For the context of **GOCAT** design, the additional Cartesian coordinate system definitions

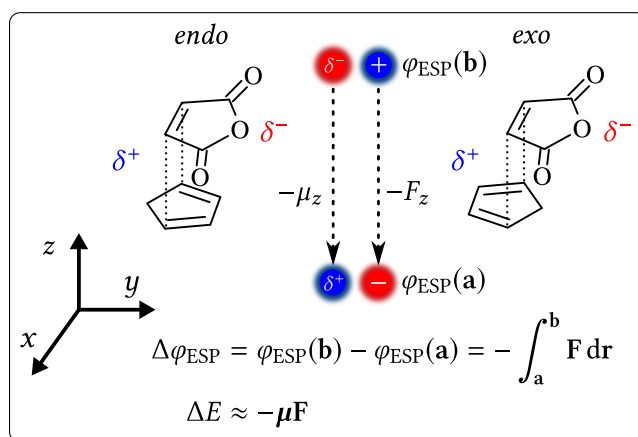


Fig. 7.1: **DA** reaction of cyclopentadiene and maleic anhydride including coordinate system definitions and both the $\Delta\varphi_{\text{ESP}}$ and \mathbf{F}_{EF} sign conventions.

are needed. In particular, the z -axis is defined for each *separate* structure of the reaction in the *molecular* frame—not in absolute laboratory frame—and points into the direction of the two newly created σ -bonds in the one-step concerted mechanism between the four involved C atoms. This is the aforementioned “reaction axis” along which the electronic reorganization (mainly) takes place, *i.e.*, along which the bonds are formed. An impact along this direction, more specifically, along the negative direction in the definitions here, can be conceptualized as field-augmented TS stabilization of ionic structures facilitating the forming of these bonds.^[268,459,463,477] The y -direction is aligned mostly along the anhydride while the x -direction is orthogonal to the vertical mirror plane of all frames, showing C_s symmetry. Also illustrated is a possible EF pointing along the (negative) of z , $-F_z$, as reaction field mostly aligned with the molecular dipole moment that is illustrated here as projected onto the same axis, $-\mu_z$. Notice the parallel alignment of the two vectors for a *stabilization* because of the inverse sign convention. That is, EF points from “plus” to “minus” and *vice versa* for the dipole moment.⁵

From the outset of this GOCAT optimization, the expectations are hence very clearly set in the following.^[268,459,463] By an increase of the field along the reaction axis (negative of z), the TS will be highly stabilized by favoring highly polarized to charged species of this structure, leading to the barrier decrease. Reaching extreme fields, a mechanistic change to a very asynchronous but still concerted mechanism is expected and, eventually, a two-step mechanism with a zwitterionic intermediate. This ionic intermediate is the result of forming first only one C–C-bond with another subsequent TS for the second C–C-bond (in Ref. [463] with help of an additional solvent, CH_2Cl_2). The C_s symmetry of the reaction is broken in this intermediate since only one bond is formed to reach the highly polarized (*i.e.*, zwitterionic) state with formally a positive charge at the cyclopentadiene part and a negative charge at the maleic anhydride moiety. Interestingly, this development of the mechanism from the symmetric concerted one to the two-step one having an ionic intermediate at high field strengths takes place *gradually* with increase of the field. Consequently, this leads to the asymmetric concerted mechanism in-between. Besides, the stereoselective outcome can be controlled by subtle different barrier decreases perpendicular to this directions (y) whereas a field orthogonal to the present mirror plane leads to no barrier effect at all (x).

7.3 Methodology

Calculations: For this DA reaction, multiple settings were examined that included $N_{\text{Ch}} \in \{10, 20, 81\}$ and always enforced charge neutrality of the GOCAT, $\sum_i q_i = 0$. The former two GOCAT sizes were optimized onto a vdW surface of the complete DA reaction path, fully aligned and discretized into 18 frames in this case. Furthermore for $N_{\text{Ch}} = 20$ in one additional setting, the minimal distance enforced between the charges i and j was relaxed to $r_{ij} \geq r_{\text{min}} = 0.1 \text{ \AA}$ instead of the usual $r_{ij} \geq r_{\text{min}} = 1.0 \text{ \AA}$ for even more flexibility of that

⁵ These directions are differently chosen from most of the literature,^[268,459,463] following the usual physical conventions and not the (outlier) ones of specific program packages the authors used. Hence, for comparison, signs and directions must be checked meticulously.

GOCAT model, *i.e.*, an even bigger search space with more possible inhomogeneous field effects. By contrast, the biggest sized model with $N_{\text{Ch}} = 81$ is bound to a sphere with a radius of $r = 7.5 \text{ \AA}$ around the geometric center of all atoms in all frames. In this model, an inter-charge distance of $r_{ij} \geq r_{\text{min}} = 3.0 \text{ \AA}$ is enforced, and this model is intended to deliver a more homogeneous field.⁶ This is due to the fact that spherical **GOCATs** have a significantly higher distance, of at least about 4 \AA , to all embedded reaction atoms and thus cannot have such a *local* impact by addressing almost single atoms as in the **vdW** case. This is also a severe advantage for the extension to be studied in this Section for the **GOCAT** optimization, namely the *adaptive* (or non-vertical) fitness function with full relaxations of all frames via the **NEB** algorithm (*cf.* Section 2.5.1.3 and Algorithm 3.2). Hence, the fitness function used was either the same as already used in Section 6.2 and described in Algorithm 3.1 (*static* or *vertical* mode) or the *adaptive* one.

However, in contrast to simpler reactions with respect to the expected electrocatalytical impact as, *e.g.*, the Menshutkin reaction (*cf.* Section 6.3.3), the **DA** reaction shows both a changing molecular dipole moment direction and magnitude during the reaction. Thus, the effects will be more subtle here and not that very well and simply linearly correlated as described for the former case. Additionally, due to the same subtleties, the fitness function for **DA** was used in two different modes that directly translates to two qualitatively different **GOCAT** optimizations: Either the **R** frame is enforced to be stabilized, $\Delta E_{\text{R}} \leq 0 \text{ kcal mol}^{-1}$, by using an additional penalty that deteriorates worse candidate solutions during the search, or without such a penalty. The latter was the case for the Menshutkin investigations (Chapter 6), for instance, too. But for the Menshutkin reaction, no embedding was intended or anticipated to ever destabilize the educt structures, which conversely is true and even exploited by the **GA** for this **DA** reaction. These two settings for the fitness function were already described in Section 3.6.1 on p. 86. In total, this leads to 18 settings, including both *exo* and *endo* cases, of up to 6000 individuals each, again consisting of multiple separate **GA** runs for $2\text{--}3 \cdot 10^6$ iterations that were locally optimized only after the **GA** and then statistically evaluated using the procedures described in Section 2.6 and used as well in Section 6.2. Though, because of the final extent of the database of all settings, illustrations of all these are outside the scope of this Thesis. Instead, the results and discussions below rather focus on the important findings that could be made with help of selected examples. All calculations shown here are again using the **PM7**^[206] semi-empirical Hamiltonian as implemented in **MOPAC**.^[226]

Moreover, to extract the (literature-known) simple catalytic trends of completely uniform electric fields along any of the three Cartesian axes defined, **GOCATs** constituting “parallel plate capacitors” were evaluated. Here, $N_{\text{Ch}} = 2 \cdot 30 \cdot 30 = 1800$ charges defined on a uniform grid of a quadratic area of $20 \cdot 20 \text{ \AA}^2$ and a distance of 20 \AA between the plates from the center of the reaction were sampled by a packing operator without any further subsequent **GA** optimizations. These are then analyzed consistently using the same methodology as for all the other (**GA**-optimized) **GOCAT** models. Here, both fields in all three Cartesian

⁶ The size was roughly set to build up a uniform embedding on the sphere while still Cartesian coordinates are also optimized that will thus also vary between the **GOCATs**; all charges will sit on the sphere, of course, but the absolute positioning is different in order to use the usual sweden operator setting as is.

directions were generated that led to either inhibition or catalysis of the DA reaction. For each separate plate **GOCAT**, the charge value, q , is the same for each charge within a plate and of opposing sign between the two plates and was successively changed to create increasing field strengths, from small ones up to very extreme ones.

Furthermore, some “measurement” descriptors for evaluating the $\Delta\varphi_{\text{ESP}}$ values are illustrated in Fig. 7.2 on the following page. Aligned with the axes, e.g., the descriptor $\Delta\varphi_{\text{ESP-}z}$ was calculated which is the difference of ESP (voltage or line integral) between two points on the z -axis. Additionally, all types of averages were computed, such as $\Delta\varphi_{\text{ESP-}zx}$, where all such differences were averaged along the x -axis while still the difference along z is measured by this descriptor. Accordingly, the descriptor “ z -pl” is the one which is both averaged along x as well as y and stands for “ z -plane”. As the fields within **GOCATs** are highly inhomogeneous (or non-uniform), this step is needed in order to be able to state the observed trends for the catalysis mechanism more clearly. For the uniform fields, all the averages are identical, of course. Such Cartesian axis descriptors were created on the minimal bounding box (as best seen in Fig. 7.2(a)). Additionally, differences at the atoms themselves were calculated, i.e., line integrals along bond axes or multiple bonds, as, e.g., the most important “3m1-2m4” descriptor. It measures the symmetrized difference of $\Delta\varphi_{\text{ESP}}$ between the four C-atoms that are involved in the bond creations (cf. Fig. 7.2(f) on the next page for the atom labels). For example, the corresponding symmetrized ESP difference for the TS frame along this descriptor is calculated as⁷

$$\Delta\varphi_{\text{ESP-TS-3m1-2m4}} = \frac{1}{2} [(\varphi_{\text{ESP-TS-C3}} - \varphi_{\text{ESP-TS-C1}}) + (\varphi_{\text{ESP-TS-C2}} - \varphi_{\text{ESP-TS-C4}})]. \quad (7.1)$$

Note that a positive $\Delta\varphi_{\text{ESP}}$ maps to expected catalytic fields with a higher potential at the anhydride side, and this leads to fields, $\mathbf{F}_{\text{EF}} = -\nabla\varphi_{\text{ESP}}$, pointing in the negative z -direction (as shown in Fig. 7.1 on p. 180). At the end, about 100 different ESP/EF and about 60 other features were analyzed, such as dipole projections, estimations with Eq. (6.2) on p. 171, etc., besides the raw (not “engineered”) features such as energies, gradients, the fitness, etc.

Electric Field Projections: Electric fields can always be derived as conservative vector field from the scalar potential function, φ_{ESP} (cf. Section 2.3). However, as in the cited literature,^[122–124,126,127,455] we are also again interested in the *local* field along a bond; this was also calculated in Section 6.3.3. To these ends, other authors compute either one single field in the middle of the interesting bonds, in our case for instance C1 and C3, or generally A and B , and project this field onto the normalized direction vector, $\hat{\mathbf{r}}_{AB}$. Or they compute an averaged version $0.5(\mathbf{F}_B + \mathbf{F}_A)$ and project that onto the bond direction. With the inhomogeneous fields that highly optimized **GOCATs** often generate in the present case, this field projection is still too erratic such that an even more averaged version is used

⁷ In vdW models, the **GOCAT** charges can be very near or even within the bounding box of the frame such that the atom-based descriptors are more meaningful in this case.

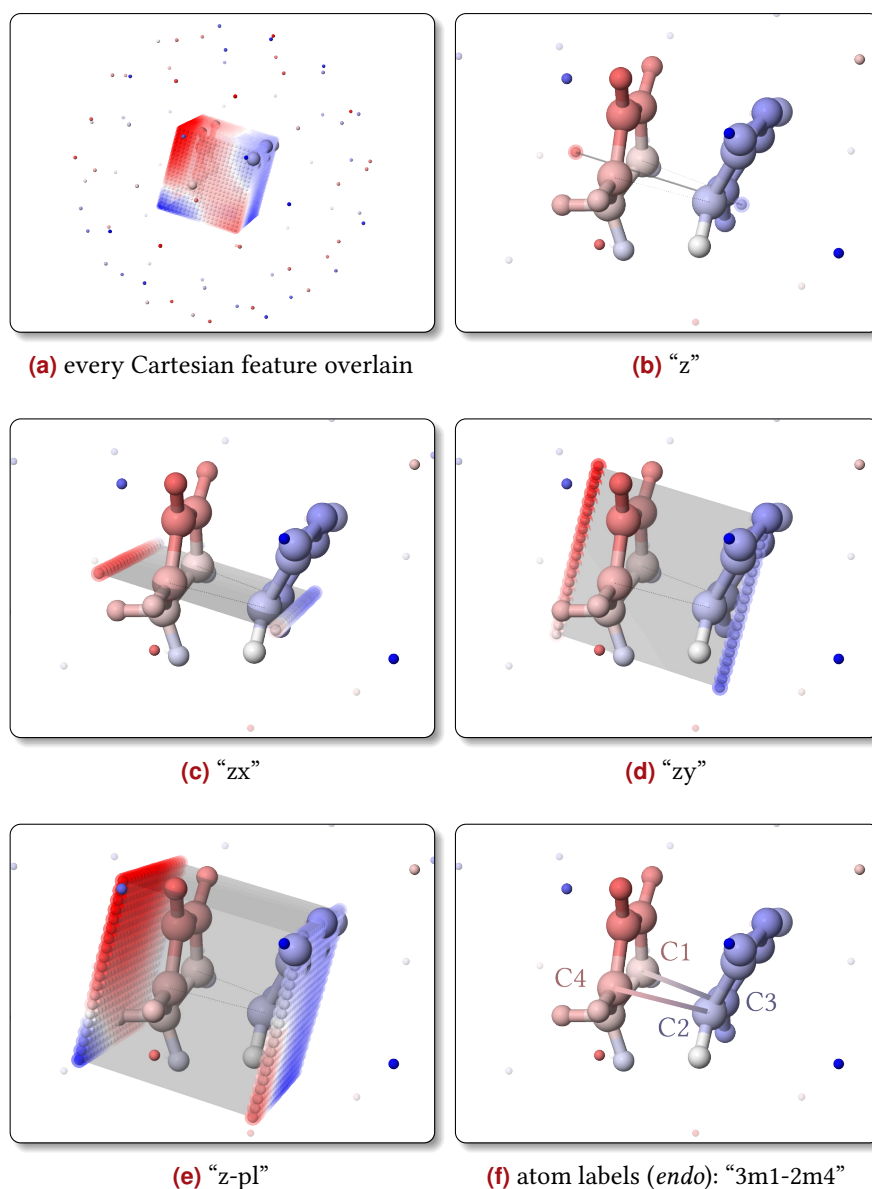


Fig. 7.2: Some EF characterizations (features) are shown, based on the Cartesian coordinate system as illustrated in Fig. 7.1. ESP values were calculated at the marked positions placed on a minimal bounding box of the frames aligned with the coordinate system. Then the $\Delta\varphi_{\text{ESP}}$ values and the mean EF were calculated between those: “z” terms the feature regarding the z-axis, “zx” the feature additionally averaged along the x-axis, but as difference along z, etc., “3m1-2m4” is the difference between the attacking/attacked atoms creating the new C–C-bonds. Similar to “z”, “zx”, ..., “z-pl”, there also exist “y”, “yz”, etc. (and hundreds of other features).

as

$$\bar{\mathbf{F}}_{\text{EF},AB} = -\frac{\Delta\varphi_{\text{ESP}}}{\|\mathbf{r}_{AB}\|} = \frac{1}{\|\mathbf{r}_{AB}\|} \int_{\mathbf{r}_A}^{\mathbf{r}_B} \mathbf{F}_{\text{EF}}(\mathbf{r}) \, d\mathbf{r} \quad (7.2)$$

$$= \frac{1}{\|\mathbf{r}_{AB}\|} \lim_{n \rightarrow \infty} \sum_i^n \mathbf{F}_{\text{EF}}(\mathbf{r}_i) \Delta\mathbf{r}_i \quad (7.3)$$

$$= \frac{\mathbf{r}_{AB}}{\|\mathbf{r}_{AB}\|} \lim_{n \rightarrow \infty} \sum_i^n \frac{\mathbf{F}_{\text{EF}}(\mathbf{r}_i)}{n}, \text{ with equidistant steps } \Delta\mathbf{r}_i = \frac{\mathbf{r}_{AB}}{n} \quad (7.4)$$

$$= \hat{\mathbf{r}}_{AB} \cdot \bar{\mathbf{F}}_{\text{EF}}. \quad (7.5)$$

Eq. (7.5) might simply emphasize that the $\Delta\varphi_{\text{ESP}}$ values that are heavily used and discussed below simply denote the average EF, $\bar{F}_{\text{EF},AB}$, where we will drop the bar over the symbol from now on. This is the averaged field between \mathbf{r}_A and \mathbf{r}_B , for instance between two C-atoms, C1 & C3, as negative ESP difference, divided by the Euclidean distance between the points. This is illustrated by the integral definition of the line integral in Eq. (7.4); in other studies,^[123,124,126,127] a numerical quadrature of $n = 2$ (at the end-points, A or B) or $n = 1$ (in the middle) was used.⁸

Evaluations: For the further evaluation methodology, such as the duplicate detection and the subsumption in chunks of representative data, *i.e.*, the unsupervised ML, the descriptors used for describing the electrostatic GOCATs (ESP difference vector norm, BOB representation, etc.), see both the general introduction in Section 2.6 and the depictions in the publication (Section 6.2 on p. 132). The same notation from the publication is used for the HC clustering again, *e.g.*, r0-c12-n324 that denotes the cluster number 12, includes $N_{\text{GOCAT}} = 324$ candidate solutions and contains as best rank (lowest fitness) GOCAT, the r0.

7.4 Results

In the following, we will focus on the *endo* DA reaction, which is the less thermodynamically stable, but faster reaction. *Static* calculations are the ones trying to “vertically” stabilize the pre-computed NEB frames in order to reach a catalytic effect and are given first (Section 7.4.1), whereas *adaptive* calculations are the ones starting also with the gas phase path but progressively reaching relaxed reaction paths within the evolving GOCATs that, at the end, might then change the mechanism of the concerted symmetric one-step one to something new (Section 7.4.2). Some further comparisons, also including the uniform fields, follow in Section 7.4.3. Most findings are already discussed right away, but some concluding remarks follow in Section 7.5.

7.4.1 Static Globally Optimal Catalysts

Spherical GOCATs: Starting with the static spherical GOCATs, some smaller effects on the reaction energy profiles in a more homogeneous embedding are shown in Fig. 7.3 on the next page. Two clusters after HC, which is illustrated in the Appendix (Figs. A.8 and A.10 on pp. 273ff.), based on the usual symmetrized ESP difference vectors are presented here. Cluster c12 with $N_{\text{GOCAT}} = 324$ individuals and the best rank found, r0 (using the notation “r0-c12-n324”), shows a barrier decrease of $\Delta\Delta E^\ddagger \approx -4 \text{ kcal mol}^{-1}$, starting at the reference barrier with no field, $\Delta E^\ddagger = 20.08 \text{ kcal mol}^{-1}$, and reaching thus a final barrier of $\Delta E^\ddagger = 16.23 \text{ kcal mol}^{-1}$. This barrier decrease is similarly observed in the cluster mean that is represented by its **nearest neighbor (NN)** with $\Delta E^\ddagger = 16.81 \text{ kcal mol}^{-1}$. The cluster c3,

⁸ $\Delta\varphi_{\text{ESP}}$ and $\bar{F}_{\text{EF},AB}$ are (almost always) perfectly linearly correlated except for average fields over differing (unsymmetric) bond distances in the *adaptive* version; in the scatter plots below, both descriptors could simply be exchanged. The fields, though, were often measured and/or computed in the other literature studies and thus EFs are also used here for comparison, when needed.

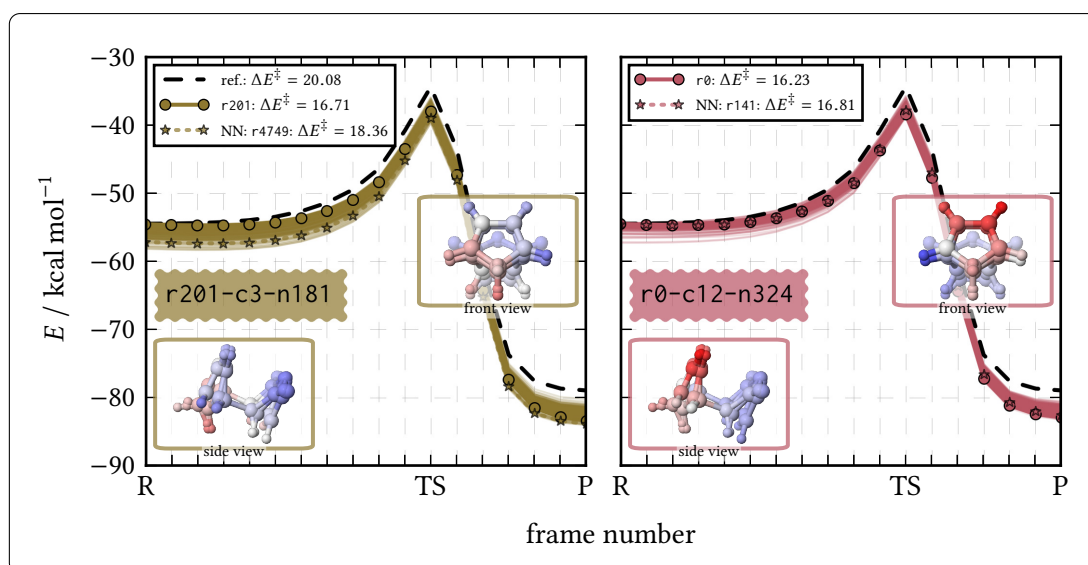


Fig. 7.3: $N_{\text{Ch}} = 81$ (sphere, static): Reaction energy profiles of **GOCATs** on a sphere for the **DA** reaction for two selected **HC** clusters: c12 with the best **GOCAT** (right panel), and another one, c3 (left panel), with the highest correlations found with respect to the z -plane. Barriers, ΔE^\ddagger , are given in kcal mol^{-1} . A corresponding dendrogram for the **HC** and an **MDS** plot are given in the Appendix in Figs. A.8 and A.10 on p. 273 and on p. 275, respectively. See the main text for details.

which contains r201, is illustrated, too, in order to discuss the problem of erratic (hidden) effects with respect to the correlated electrostatic properties with the barrier decrease and is also similar to c12 regarding the energetic properties. What is more, also all non-identical $N_{\text{GOCAT}} = 4070$ **GOCATs** in the final population without clustering all share similar reaction energy profiles as these two selected clusters if they were plotted together (not shown).

As this is the fitness function with an additional penalty, if the reactant is destabilized, $\Delta E_{\text{R}} > 0$, with $\Delta E_{\{\text{R,TS}\}} = E_{\{\text{R,TS}\}} - E_{\{\text{R,TS}\},\text{ref}}$, we see overall stabilizations of both **R** and, more importantly, of **TS**. Of course, always $\Delta E_{\text{TS}} < \Delta E_{\text{R}}$ must hold (with negative stabilization energies) in order to generate the barrier decrease. Without enforcing $\Delta E_{\text{R}} \leq 0$, there is a giant search space with candidate solutions exploiting the **R** destabilization with less (or no) **TSS** at all. One such example is given in the Appendix (Figs. A.13 and A.14 on p. 277).

From the inset structures in Fig. 7.3, some **ESP** difference in the z -direction is already apparent, but also some variance in the other directions. These structures are the stationary structures superposed and embedded by the *mean* **ESP** of the clusters; the **ESP** features will be discussed below, again.

In Fig. 7.4 on the next page, the best spherical **GOCAT** found, r0, is shown. First, in Fig. 7.4(a), all $N_{\text{Ch}} = 81$ charges constituting the surrounding sphere around the three stationary structures can be seen. By using the sphere, a rather uniform field is created, as expected, but at the same time, a giant search space is present that originates charges, *e.g.*, on one half or part of the sphere, that by superposition “work together” to influence the core structures. An additional observation more striking for spheres is that the charges in the Cartesian domain can be interpreted less easily than in the **vdW** case (shown below).

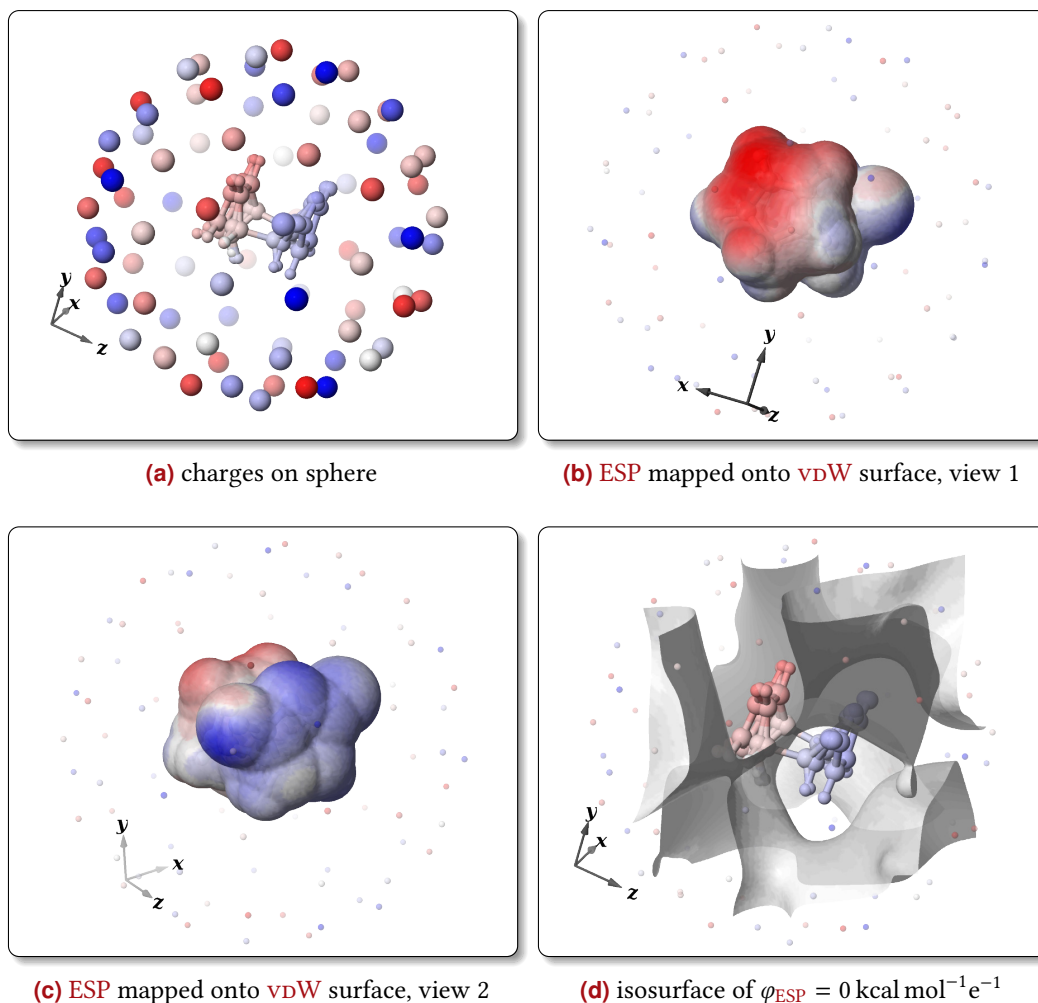


Fig. 7.4: $N_{\text{Ch}} = 81$ (sphere, static): Best rank, r_0 , found, showing the charges, φ_{ESP} mapped onto the vdW surface of the three stationary frames (R, TS, P) and the nodal surface of $\varphi_{\text{ESP}} = 0 \text{ kcal mol}^{-1} \text{ e}^{-1}$. Atoms, charges and the vdW surface are colored from red to blue for $q_i \in [-1, +1] \text{ e}$ and $\varphi_{\text{ESP}} \in [-35.0, 35.0] \text{ kcal mol}^{-1} \text{ e}^{-1}$. The same coordinate system that was introduced in Fig. 7.1 is shown.

Palpably, there are charges at different places that are almost screened by the neighbors such that no sphere with one strictly negative (red) and one strictly positive (blue) half with continuous progression in-between could be found at all. Instead, a more seemingly random blend of those charge values all around is frequent. This is due to the (already often stated) fact that the charge values and places are very probably underdetermined in order to create a very similar ESP (as discussed in, e.g., Section 2.1.1.2). Hence, multiple different Cartesian realizations, corresponding to local optima in the search space, of the same ESP exist within *one* cluster, besides the even more multiple local optima in different ESP domains. Nevertheless, as seen in Figs. 7.4(b)–(c), the final ESP has a strong z -axis difference: A negative ESP (red) at the diene and a positive one (blue) at the anhydride dienophile site can be seen. Moreover, the nodal surface of $\varphi_{\text{ESP}} = 0 \text{ kcal mol}^{-1} \text{ e}^{-1}$ showing the sign change, *i.e.*, the embedding that cuts between negative and positive ESP domains,

is pictured in Fig. 7.4(d).⁹ This again underpins the aforementioned findings of both the z -axis aligned potential (cutting in the middle of the C–C-bonds, but not fully orthogonal to these newly created bonds, *vide infra*), and some random search space artifacts as there is not just *one* plane-like nodal surface, but a very complex embedding distant from the molecular center of the frames and hence from the atoms.

These more random charges that are partially screened might point to *domino convergence* of the GA. This can be observed when there are features of the candidate solutions that contribute significantly differently to the fitness such that the more important ones converge and the less important ones get a too big freedom. Hence, the less relevant features are, if at all, fine-tuned at the end, while the main basin of attraction precedence was dominated by the more important ones.^[29] In the spherical GOCATs setting, as one could imagine, there should in principle certainly be solutions that are less complex, *i.e.*, without charges of opposite sign sitting side by side, with a similar or even better EF with regard to catalysis.

Next, in Fig. 7.5 on the following page some pair-wise relationships of **R** stabilization, $\Delta E_{\mathbf{R}}$, **TS** stabilization, $\Delta E_{\mathbf{TS}}$, the barrier, ΔE^{\ddagger} , and the **ESP** difference in z -directions, averaged along x and y (“ z -pl”, Fig. 7.2(e)), are plotted for the complete population. Apparently, the overall database shows a moderately high linear correlation of the barrier, ΔE^{\ddagger} , with respect to $\Delta\varphi_{\mathbf{ESP-TS-p-pl}}$, with a total of $R^2 = 0.546$ variance explained by the latter. This trend is amplified further by the colormap since the best ranks (deep red) show the lowest barriers and a progression to the worse ones (yellow) along the trend can be seen: By increasing $\Delta\varphi_{\mathbf{ESP-TS-z-pl}}$, the catalysis is increased.

However, the picture is quite more complicated. By looking at $\Delta E_{\mathbf{R}}$, we see the artificial truncation of the data for $\Delta E_{\mathbf{R}} > 0$ due to the fitness penalties. However, this fitness function setting is needed here as otherwise the population gets dominated by the **R**-destabilizing solutions. There seems to be an intrinsic trend to destabilize the **R** frame, which actually would speak for the “anticipation of a **TS** electrostatic need” in disfavor of that of **R**, *cf.* Section 6.3.3.1, but even this picture is a bit more involved.

Looking at $\Delta E_{\mathbf{TS}}$, we even see no or only a vanishingly small correlation at all ($R^2 = 0.012$). Actually, the p -values were always calculated alongside, but only annotated if they are significantly bigger than zero.¹⁰ With $p = 8.6 \cdot 10^{-13}$, this is essentially still zero such that

⁹ Note that the sphere and the enforced neutrality lead to the relevance of that nodal surface generating GOCATs with a clear zero **ESP** somewhere in the center. All potentials could be shifted by a constant as the fields stayed the same.

¹⁰ For the statistical tests, the following generally holds: The null hypothesis, H_0 , is the independence of the predictor (abscissa, x) and the response (ordinate, y) variable, *i.e.*, a slope of zero or no association, with the probability, p , obtaining a result at least as extreme as the current one by chance while assuming that H_0 is true. Hence, if the p -value is very small, we can reject the null hypothesis in favor of the alternative hypothesis, H_1 (that is not itself *proven* in this way), and state a linear correlation between the predictor and the response variable (in a two-sided t -test). Then, in the long run, we make an error with the probability p (type one error) that H_0 (no association) is true but was erroneously rejected. Note that in order not to over- or underestimate such probabilities, a couple of assumptions have to be true, *i.e.*, linearity of the scatter points, normality of the residuals scattering around zero (though, small deviations are fine, if the sample size is big enough because of the central limit theorem), homoscedasticity (homogeneity of variance) and independence of errors. Notice that some of these assumptions are violated in some cases when, *e.g.*, there is a distinct non-ellipsoidal shape of the scatter plot and, *e.g.*, also the GA uses some “timeline” (propagation of older results/information) that cannot be assumed to lead to independence of the data points. Therefore, the annotated p -values are only reliable in the scatter plots where those assumptions hold. But because of the size

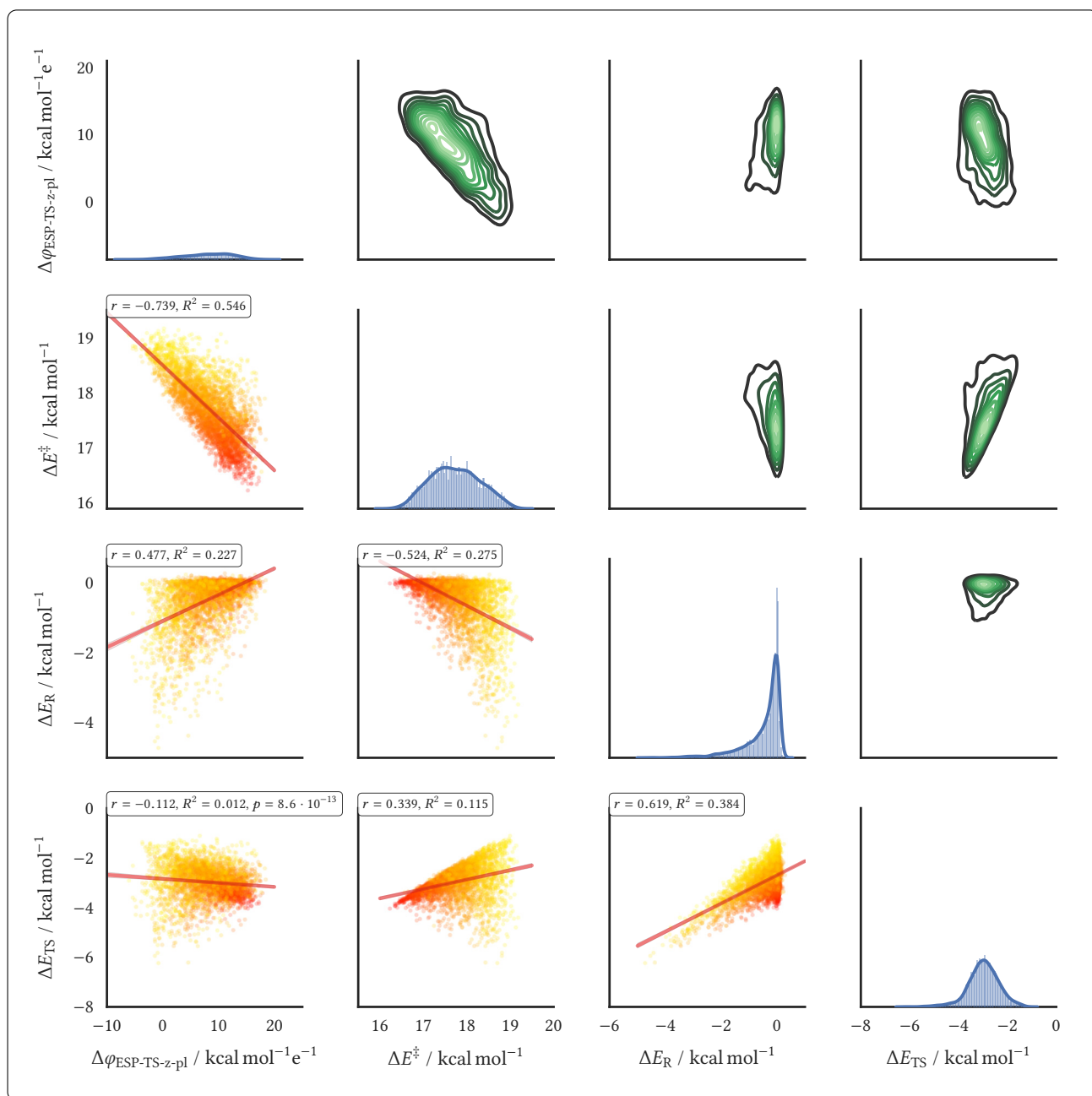


Fig. 7.5: $N_{\text{Ch}} = 81$ (sphere, static): Pairwise relationships of all GOCATs ($N_{\text{GOCAT}} = 4070$) without any clustering. In the lower triangle, correlation (scatter) plots are shown with regression lines and a colormap between red/yellow for low/high fitness values. On the diagonal, a histogram (smoothed with a kernel density estimate on top) is shown and in the upper triangle just the latter kernel density estimate for the corresponding 2D relations; the kernel density estimate is one non-parametric way for estimating the probability density function of the random variables, which is used here mainly for illustration purposes.^[354] $\Delta\phi_{\text{ESP-TS-z-pl}}$, ΔE^\ddagger , ΔE_{R} and ΔE_{TS} are the ESP difference between the z-planes (the descriptor shown in Fig. 7.2(e)), the resulting reaction barrier, the R stabilization (negative energy differences) and TS stabilization, respectively.

the barrier decrease itself, ΔE^\ddagger , is not correlated with TSS (alone) in the whole population. That is, the TS frame is just slightly stronger stabilized (with a small negative r) along the (itself highly averaged) z -plane ESP difference. As will be discussed below, just a fraction of the population follows a more clear TSS stabilization mechanism, but the overall population contains a lot of “noise”. That is, other mechanisms are superposed such as R destabilization as well as inhomogeneous fields at specific atoms, which cannot be explicated by this z -plane.

Looking at ΔE_{TS} vs. ΔE^\ddagger , we see a highly inhomogeneously varying scattering¹¹ with one boundary being exactly the TSS. This means that the best and lowest barriers are bounded by the TSS mechanism, but the other GOCATs with less ΔE^\ddagger follow other trends. By the relation $\Delta E_{\text{static}}^\ddagger = \Delta E_{\text{TS}} - \Delta E_{\text{R}} + \Delta E_{\text{ref}}^\ddagger$,¹² a complete linear (positive) association between $\Delta E_{\text{static}}^\ddagger$ and ΔE_{TS} is expected when ΔE_{R} is zero, which is not the case and creates the spread by the variance of both stabilizations.

With a positive linear correlation between ΔE_{TS} and ΔE_{R} , we see that there are GOCATs that have stabilized both frames at the same time, but by different amounts in order to create a lower barrier. Otherwise, there would be an overly vertically shifted or stabilized gas phase path without any *relative* energetic changes. Indeed, a big part of the population peaks slightly below $\Delta E_{\text{R}} = 0$ (including the best ranks), *i.e.*, these GOCATs work by TSS alone.

For following the last line of thought, Fig. 7.6 on the next page shows a similar matrix correlation plot, but now only for $N_{\text{GOCAT}} = 324$ GOCATs in cluster c12 that contains the best rank, r_0 . Here, we have a lower, but still a high correlation of ΔE^\ddagger with respect to $\Delta\varphi_{\text{ESP-TS-}z\text{-pl}}$, but again we observe some noise when looking at ΔE_{TS} vs. $\Delta\varphi_{\text{ESP-TS-}z\text{-pl}}$. Interestingly, as indicated above, this cluster has a higher correlation of ΔE_{TS} with the z -plane ESP, but still *only* $R^2 = 0.111$, meaning that almost 90% of the variance in TS energies is still showing up together with other ESP influences than the ones along the z -planes alone. The highest association indeed is the “3m1-2m4” descriptor (*cf.* Fig. 7.2(f)), with $r = -0.439$, $R^2 = 0.193$ (not shown), *i.e.*, without any averaging alongside the other axes of the minimal bounding box planes. Nevertheless, ΔE^\ddagger can be traced back to ΔE_{TS} since the most GOCATs spread around $\Delta E_{\text{R}} = 0$. Yet, there is *no* simple property that was looked at, which included essentially all types of line integrals along the molecule, for all frames, that alone describes the TSS. That is, for this GOCAT model (sphere, static) there is

of the database, the p -values are then (often) already numerically zero. However, note also that this procedure (looking at correlation coefficients and hypothesis tests) was used in the background in a high-throughput screening to search for correlations of all the other descriptors of GOCATs and are annotated in any case, also when the assumptions might be violated, which can be estimated from the raw scatter points.

¹¹ This violates one of the assumptions for the linear correlation analysis and thus correlation coefficients should not be over-emphasized and the raw data should stand on its own.

¹² $\Delta E_{\text{static}}^\ddagger$ is the barrier between the gas phase TS and R frames, whereas $\Delta E^\ddagger \geq \Delta E_{\text{static}}^\ddagger$, by a small amount, is the effective (maximum) barrier between the frames as seen by the fitness function. This barrier is only slightly higher if the R frame is shifted, besides the additional penalties on small shifts of the lowest energy along the reaction path (described in Algorithm 3.1 on p. 87 and used in Table 7.1 on p. 203). This leads to a small additional noise in the scatter plots shown here but to the same general observations. Everything was also evaluated using $\Delta E_{\text{static}}^\ddagger$ instead.

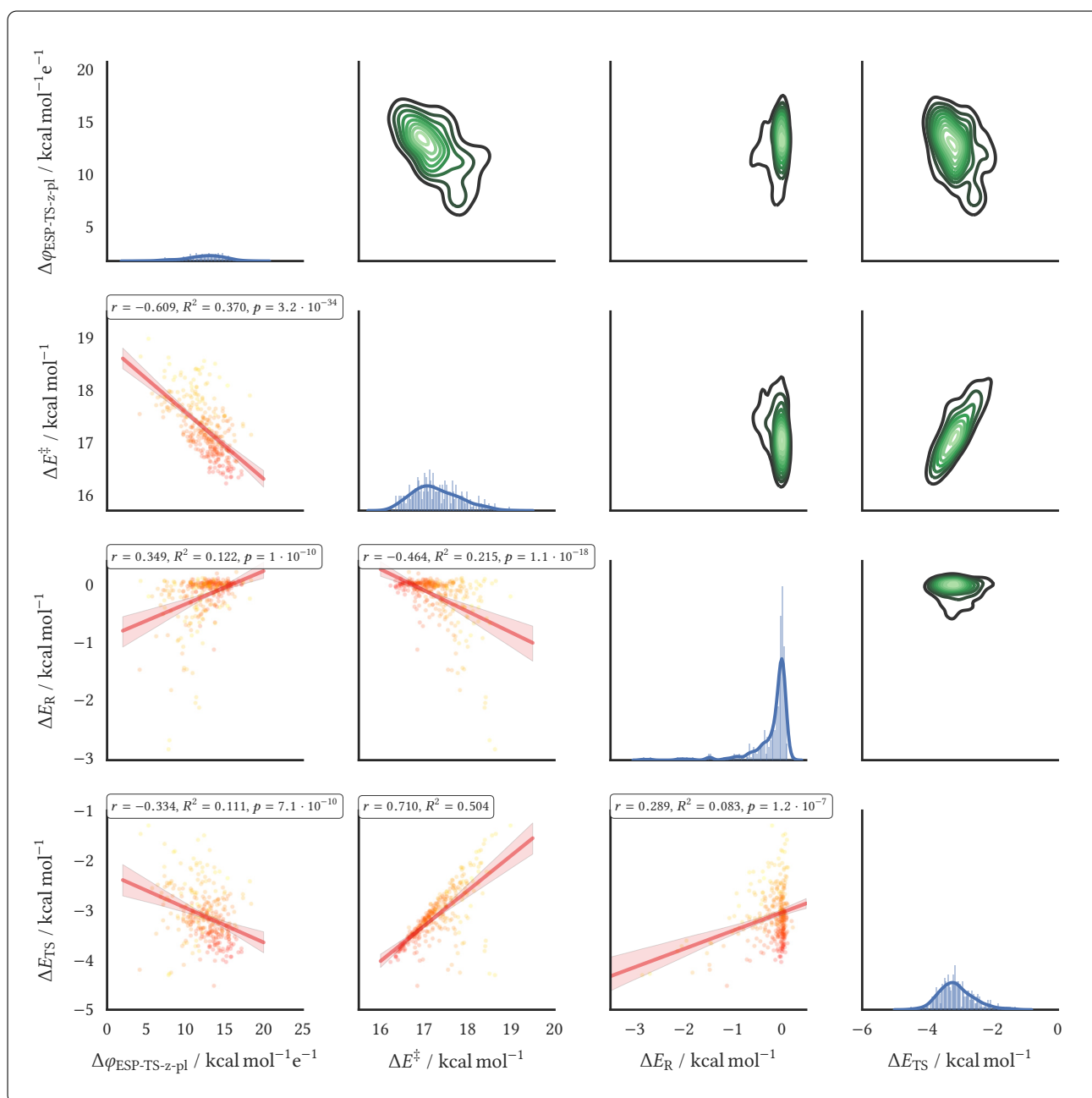


Fig. 7.6: $N_{\text{Ch}} = 81$ (sphere, static): Pairwise relationships of the $N_{\text{GOCAT}} = 324$ GOCATs of the best cluster, c12. For plotting details, see Fig. 7.5. Additionally, in the cases shown here, finite (reddish) 0.95% confidence intervals for the regression estimate around the main regression lines can be seen, which were not observable in Fig. 7.5 because of the higher statistical significance of the cases of Fig. 7.5.

already a very complex property landscape by which such subtle catalytic effects can be generated, not just by (uniform) fields along z .

By optimizing globally inhomogeneous fields for such a reaction profile manipulation, all different types of effects are possible. For instance, a vertical shift of all frames' energies by an y -field or even an x -field, see below and Fig. A.19, creating $\Delta E_{\text{TS}} < 0$, then some **R** destabilization by introducing a field along some exposed H-atoms of the **R** structure.¹³ Thus, this non-simplistic nature of the (subtle) effects showing up here is understandable, and this is indeed quite more involved than the Menshutkin reaction, which can be reduced to a single linear **ESP** trend (*cf.* Section 6.3.3). Accordingly, maybe just the best ranks should be analyzed more thoroughly because these **GOCATs** accumulate the best mechanism possible, within the fitness function defined, and show less noise of worse local optima within their corresponding clusters.

vdW-based GOCATs: Leaving the spherical **GOCATs** for a moment, the $r\theta$ of the **vdW** model with $N_{\text{Ch}} = 10$ is described next. With charges placed nearby distinct atoms, the **EF** will even be more locally varying or inhomogeneous and, consequently, the charges fulfill a more concrete “role” when not being screened by nearby charges that easily. The overall flexibility of the model increases, detectable by the minimal fitness that can be reached. As already found in Section 6.2, there is often a convergent minimum fitness value with the raise of the **GOCAT** complexity, *i.e.*, a maximum of the catalytic effect. When the flexibility or complexity of the **GOCAT** model is increased, the globally best fitness value first drops

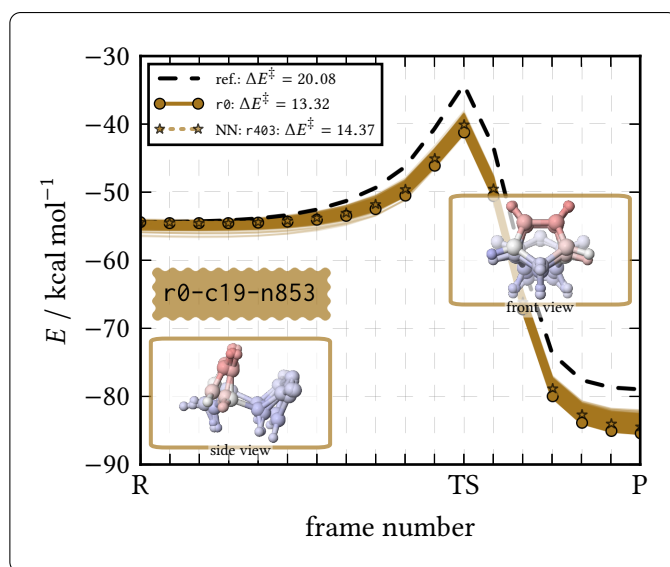


Fig. 7.7: $N_{\text{Ch}} = 10$ (**vdW**, static): Reaction energy profiles of the best cluster, c19. Barriers, ΔE^\ddagger , are given in kcal mol^{-1} . A corresponding dendrogram for the **HC** and an **MDS** plot are given in the Appendix in Figs. A.9 and A.11 on p. 274 and on p. 276, respectively. See the main text for details.

¹³This is just to be considered as an example of a qualitative argument for describing different and inhomogeneous field effects. Such arguments in a quantitative setting would need the full series of Eq. (2.67) to decompose the energetic contributions.

but reaches a plateau. After that, more complex **GOCAT** models do not lead to an enhanced catalytic effect anymore but only to more intricate surroundings for the same effect. Hence, in the case of these **vdW**-based **GOCATs** compared to the spherical ones, the reaction energy profiles show lower barriers and an amplified **TSS** with less variance at the **R** side. This is clearly visible in the reaction energy profiles in Fig. 7.7 on the previous page.

Although, for the **vdW** model, very different **GOCATs** can also be found in the same cluster with regard to the Cartesian domain (these are given in Fig. A.12), the best one, r_0 , indeed often comes out being simpler showing (approximately) a higher symmetry if the reaction itself is symmetric. This was true for the Menshutkin reaction (Figs. 6.7(a) and 6.7(b)) and is also true for the **DA** reaction. By this, **ESP** values and other properties of the best rank are often rather outliers with regard to the rest of the population (or the same cluster). Nevertheless, such distinct solutions are exactly what one intends by the global optimization when having reached a converged optimum.

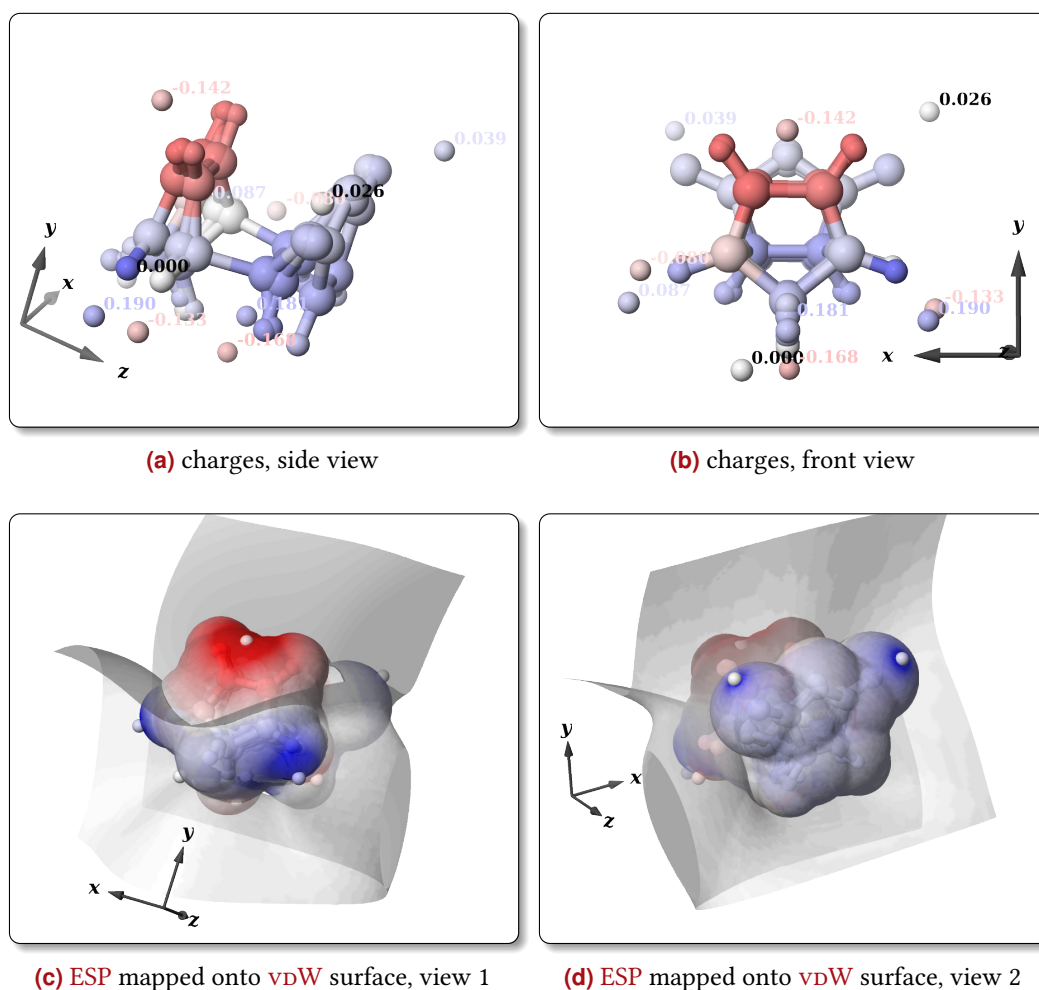
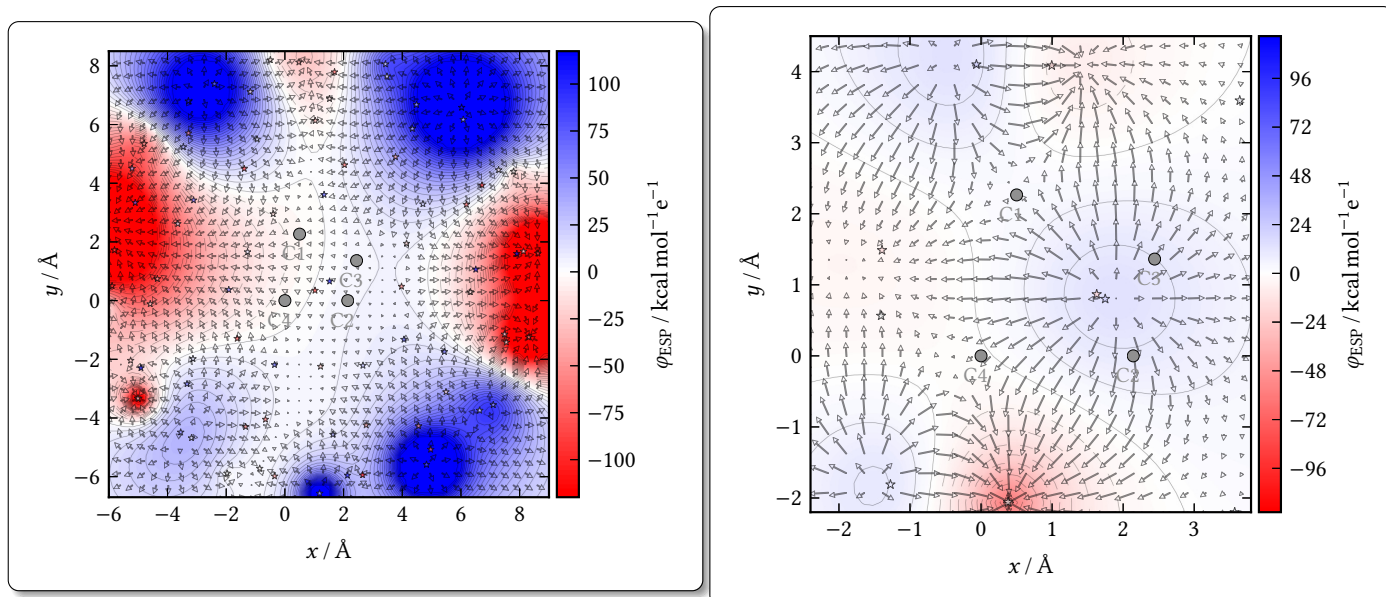


Fig. 7.8: $N_{\text{Ch}} = 10$ (**vdW**, static): Best and very symmetric rank r_0 found. Partial charges are explicitly shown in Figs. (a)–(b). φ_{ESP} values mapped onto the common **vdW** surface of the three stationary frame (**R**, **TS**, **P**) together with the nodal surface of $\varphi_{\text{ESP}} = 0 \text{ kcal mol}^{-1} \text{ e}^{-1}$ are illustrated in Figs. (c)–(d). The entities are colored from red to blue for $q_i \in [-1, +1] \text{ e}$ and $\varphi_{\text{ESP}} \in [-35.0, 35.0] \text{ kcal mol}^{-1} \text{ e}^{-1}$.



(a) $N_{\text{Ch}} = 81$ (sphere, static): r_0 , TS frame

(b) $N_{\text{Ch}} = 10$ (vdW, static): r_0 , TS frame

Fig. 7.9: $N_{\text{Ch}} \in \{10, 81\}$ (static): ESP, φ_{ESP} , and its EF as the negative gradient plotted as arrows, projected onto the plane of the four atoms (C1–4) for the TS frames of the *endo* DA reaction shown in Figs. 7.4 and 7.8. The gradients are clipped when becoming too large for visualization purposes because of the singularities at the Coulomb charge centers. Contour lines are drawn for each $\Delta\varphi_{\text{ESP}} = 5 \text{ kcal mol}^{-1} \text{ e}^{-1}$. The same colormap for each plot will be used (for comparison), also for the *adaptive* GOCATs in Fig. 7.14 below, which is the reason why such a huge range for φ_{ESP} is already used in the present cases.

This symmetric candidate solution for the vdW case is given in Fig. 7.8 on the previous page. Of the 10 charges in r_0 , one is almost zero such that only nine charges “survived” in this individual. The C_s symmetry is also almost discernible. The aforementioned positive (blue) H-atom potential (exposed atoms of R frame) can be seen and, moreover, a trend (from left to right in that Figure) along the z -axis, but also one along the y -axis (bottom to top). The nodal surface has an overall “Y”-shape and cuts through the molecule at the new C–C-bonds (actually even an “X”-shape, but the lower edge is already further away from the atoms).

In Fig. 7.9, there are the ESP scalar and EF vector fields that are given for both the r_0 GOCATs of the sphere and of the vdW model on a plane through C1–4 (*cf.* Fig. 7.2(f) on p. 184). This summarizes what was seen so far. A z -trend with contour lines cutting in the middle of C3 & C1 and C2 & C4 (Fig. 7.9(a)) and some more local, but rather symmetric fields for the vdW GOCAT are visible (Fig. 7.9(b)). Here again, not just one simple descriptor but multiple ones are working together that cannot easily be differentiated by a single linear correlation. Each separate correlation could not simply be divided into a linear combination of predictor variables. The latter are highly correlated themselves, of course, and work synergistically.

7.4.2 Adaptive Globally Optimal Catalysts

Going over to the adaptive or non-vertical **GOCAT** optimizations, the reaction energy profiles of three selected individuals that follow clearly the expected shift of the mechanism are plotted in Fig. 7.10 on the following page.¹⁴

Generally, the fitness function includes the exact same ingredients as in the last Section. However, in each evolving field, the reaction path is fully relaxed *via* the adaptive **NEB** (*cf.* Algorithm 3.2 on p. 93 in Section 3.6.3). This creates not just “vertical” energetic shifts of the profile, but the structures themselves can vary strongly. The final barriers of the **GOCATs** are significantly lower than in the static case and the accumulated path gets longer; note that in Fig. 7.10 the Euclidean distances between the discrete frames are therefore plotted on the abscissa. Also the final frame number is dependent on the adaptive **NEB** creation for ensuring enough resolution of the varying path. The longer reaction coordinate stems from a higher distance between the new **R** and **P** frames by mainly some small rotations of these within the field and—with very strong fields—by the mechanistic change, discussed below. Now, r1 shows two distinct **TSs**. Looking at the final gradient norms, the stationary points are well optimized, *i.e.*, $\|\nabla E_{\{R,TS,P\}}\| \ll 10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, in contrast to the static paths, as expected. However, with the relaxations in strong fields, and even more so if a **vdW** surface were used because of the local effects, optimizations might turn out to be either not converging at all—these would be **GOCATs** that would not survive—or to show slightly bigger gradients than in the pristine gas phase at the end.¹⁵ Note that r1 even has almost no barrier at all, *i.e.*, $\Delta E^\ddagger = 3.40 \text{ kcal mol}^{-1}$.

The detailed illustrations of these three individuals are given in Figs. 7.11 to 7.13 on pp. 197–199. Starting with r987, we still have a symmetric (synchronous) attack. The distances between the respective C atoms are $d = 2.20 \text{ \AA}$ for both pairs. In contrast, these distances were $d = 2.14 \text{ \AA}$ in the gas phase **MEP** before. Interestingly, as will be compared below again, the possible fields by **GOCATs** become more extreme because these fields will not automatically generate large gradient norms that are just penalized as done in the static case without relaxation. With an increase of the fields and a resulting stronger impact on the reaction, the main component of the field becomes better aligned with the main reaction axis, *i.e.*, this generates an even better visible field along the (negative) *z*-direction. This is clearly apparent in Figs. 7.11(c) and 7.11(d).

The r6 **GOCAT** shows already an unsymmetrical (asynchronous) concerted **TS**, but it still follows a one-step mechanism, as illustrated in Fig. 7.12 on p. 198. The distances between the respective C-atoms now are $d_{3m1} = 2.53 \text{ \AA}$ and $d_{2m4} = 2.14 \text{ \AA}$ with the former distance

¹⁴Maybe the question arises why r0 is not shown. This latter **GOCAT** is very similar to r1 in all respects, but shows a two-step mechanism slightly worse resolved due to the concrete **NEB** settings that were used. As discussed below (Section 7.5), the results can be improved by some adaptations of the **MEP** relaxation scheme in the future.

¹⁵For instance, using strict line searches in the local optimization can lead to $E = -\infty$ by fusion of atoms and charge centers (by the non-convex Coulomb potential) if the attractive potential by the rest of the system is not high enough and/or there are non-compensated gradients by the charges. Therefore, **FIRE** or **LBFSGS** with more robustness ingredients are needed, such as restarts at fluctuating directions, rescaled maximum step lengths and caching of best structures found. Usually in real chemical systems (described by **QM/MM**), there is (at least) the **vdW** repulsion leading to proper local minima on the **PES**. By using spherical **GOCATs**, the fusion is less a problem, which is the reason for utilizing them in the first place.

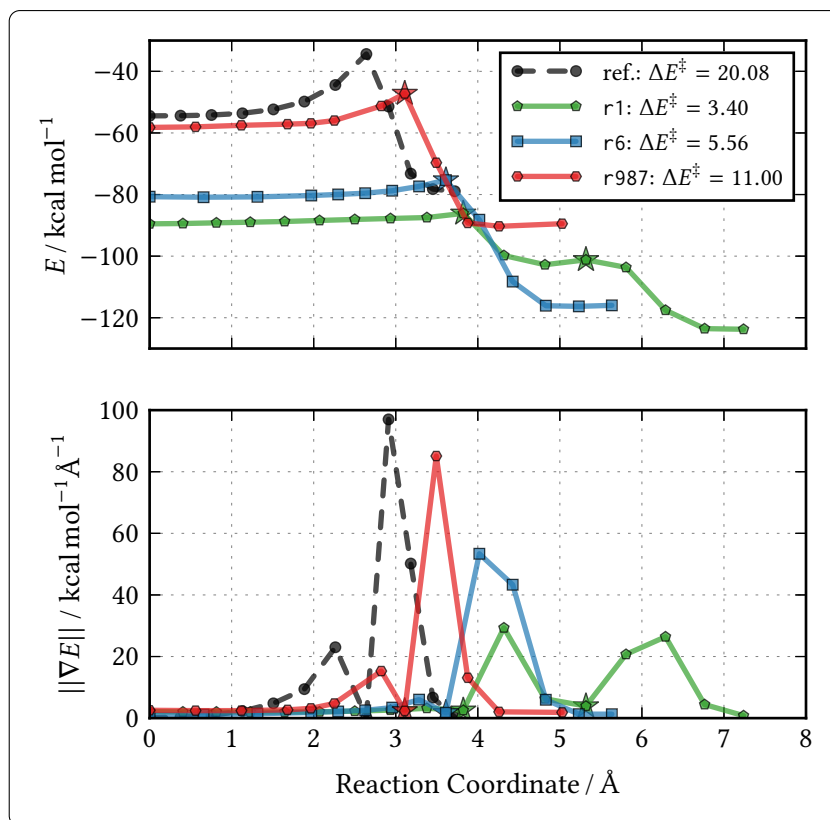


Fig. 7.10: $N_{\text{Ch}} = 81$ (sphere, adaptive): Reaction energy profiles and corresponding gradient norms of three selected *adaptive GOCATs* (r1, r6, r987) and the reference profile. Newly found **TS** frames during **GOCAT** optimization are emphasized with stars. Barriers, ΔE^\ddagger , are given in kcal mol^{-1} . The abscissa is continuous in this case since the **MEPs** can vary due to the relaxations in the **GOCATs**.

being already by $\Delta d_{\text{async}} \approx 0.40 \text{ \AA}$ longer than in the gas phase path. The field strengths are still becoming successively higher and the nodal surface simpler.

Finally, r1 shows a two-step mechanism with a zwitterionic intermediate, which is given in Fig. 7.13 on p. 199. The first **TS** creates the first C–C-bond, here between the C4 and C2 atoms with $d_{2m4} = 2.20 \text{ \AA}$, while the other C atom is still very distant with $d_{3m1} = 3.10 \text{ \AA}$. The second **TS** (with a small local minimum in between, cf. Fig. 7.10) then follows immediately with $d_{3m1} = 2.63 \text{ \AA}$. The field strength is one of the largest found so far while leading to converging relaxed **GOCATs**. These field strengths and other properties will be discussed below again, also giving some numbers then.

Similarly to Fig. 7.9, φ_{ESP} and F_{EF} fields in the plane of the four involved C atoms are given in Fig. 7.14 on p. 200. As expected and already mentioned above, the fields increase a lot and show even more distinct contour lines between, on the one hand, C3 & C1 and, on the other hand, C2 & C4, and a higher number of lines since these are equidistantly plotted in each case. Additionally, an increasing asymmetry between the bonds going over from r987 to r6 and r1 in accordance with the asymmetric **TS** they show can be stated. This is a smaller x -field that sets in here.

One possible explanation for this one could suppose is that the asynchronous and finally

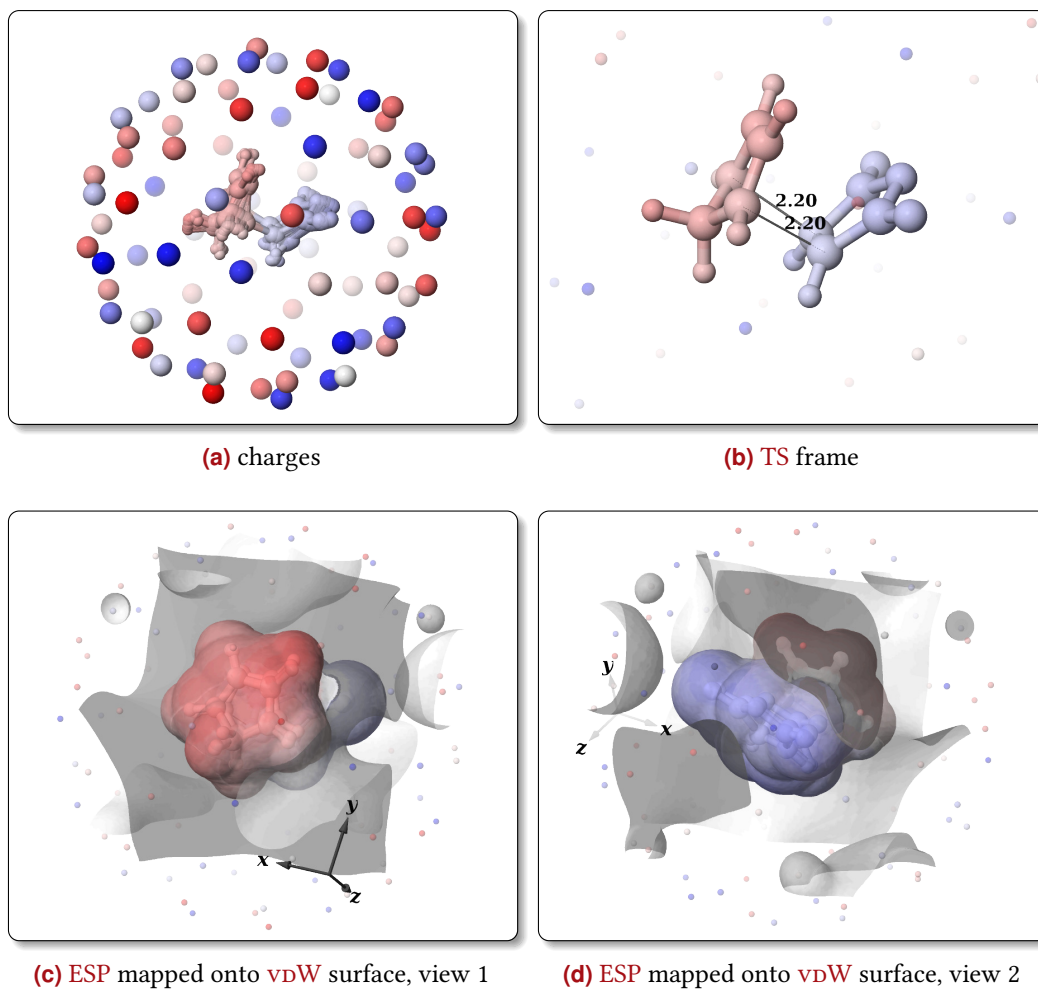


Fig. 7.11: $N_{\text{Ch}} = 81$ (sphere, adaptive): GOCAT r987 of Fig. 7.10 stabilizing a strongly catalyzed *endo* DA reaction showing (still) a synchronous concerted attack. The distances of the attacking/attacked C-atoms are given in Ångstrom in Fig. (b). φ_{ESP} values are mapped onto the common vdW surface of the three stationary frames (R, TS, P) together with the nodal surface of $\varphi_{\text{ESP}} = 0 \text{ kcal mol}^{-1} \text{ e}^{-1}$, illustrated in Figs. (c)–(d). The entities are colored from red to blue for $q_i \in [-1, +1] \text{ e}$ and $\varphi_{\text{ESP}} \in [-116.8, 116.8] \text{ kcal mol}^{-1} \text{ e}^{-1}$.

two-step mechanism is favored by introducing some x -field variance, similar to what is known from other DA reactions with already *asymmetric* electron pushing and electron withdrawing functional groups at the reaction partners. Such an additional polarization along x could also be induced by the field, and any field in x would break the C_s -symmetry already. However, there was no significant x -field correlation found at all in the overall population (not shown here). Thus, even though such x -field components can be quite big but significantly smaller than along the z -direction, no clear relation to the catalytic effect can be stated and this could rather be a random feature that is scattered in the pool showing almost a Gaussian shape.

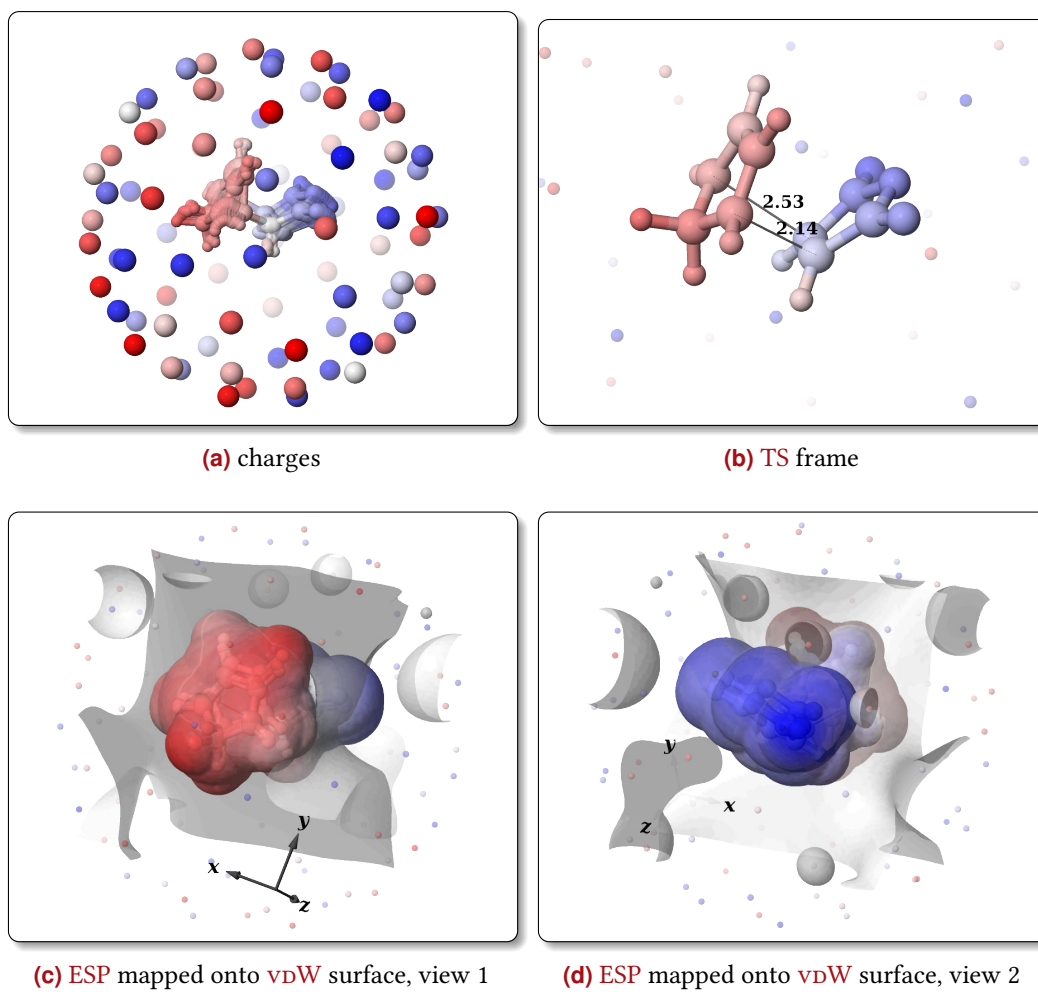


Fig. 7.12: $N_{\text{Ch}} = 81$ (sphere, adaptive): **GOCAT** r6 of Fig. 7.10 stabilizing a strongly catalyzed *endo* DA reaction showing an asynchronous concerted attack. The plotting details are the same as in Fig. 7.11.

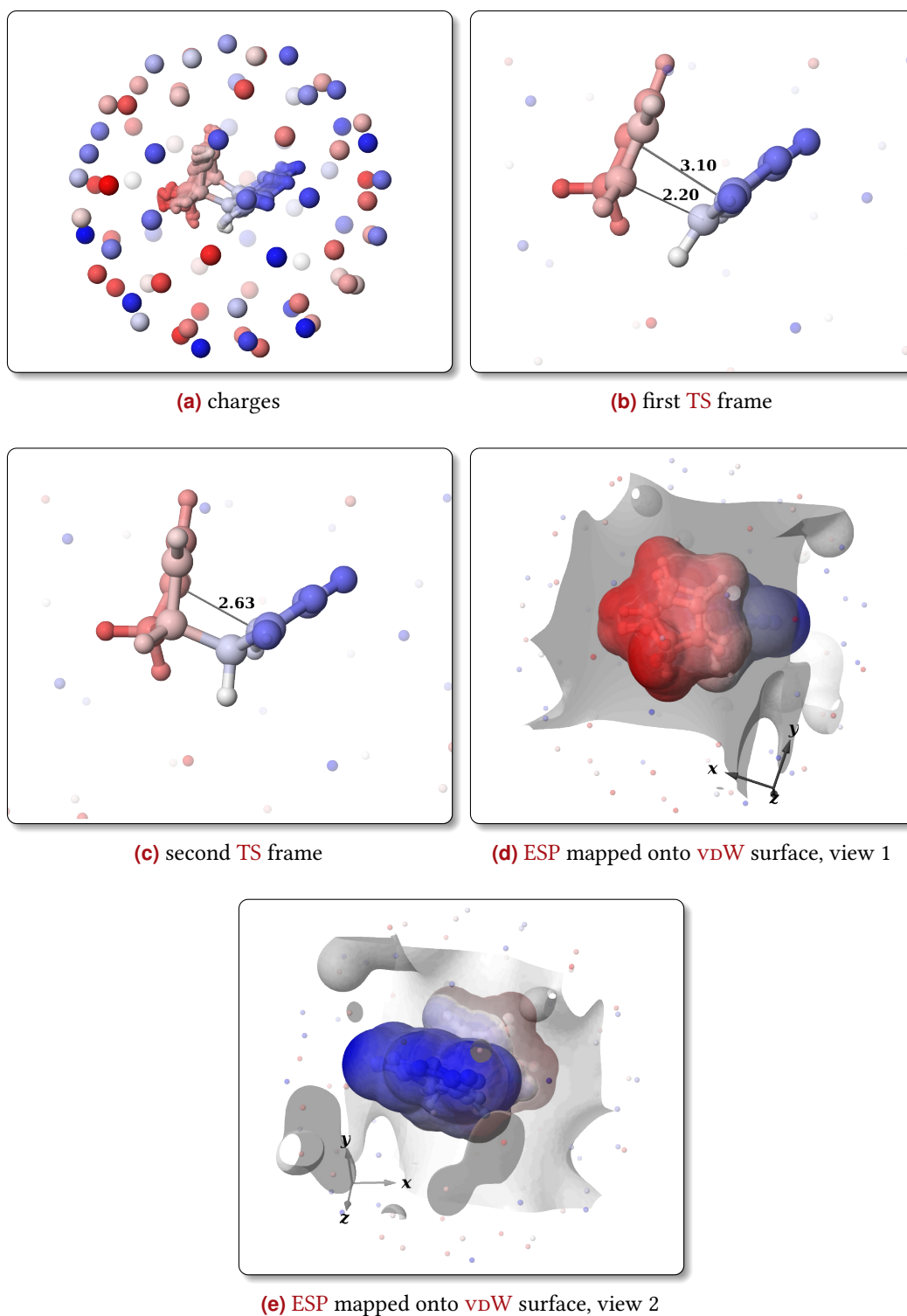
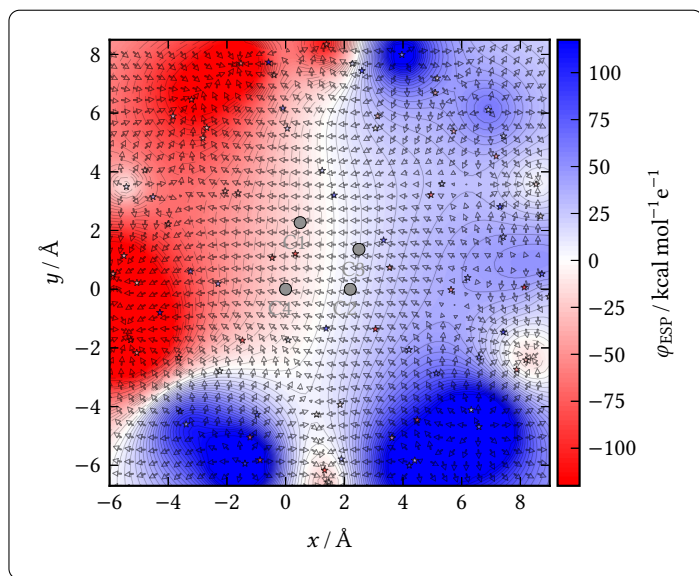
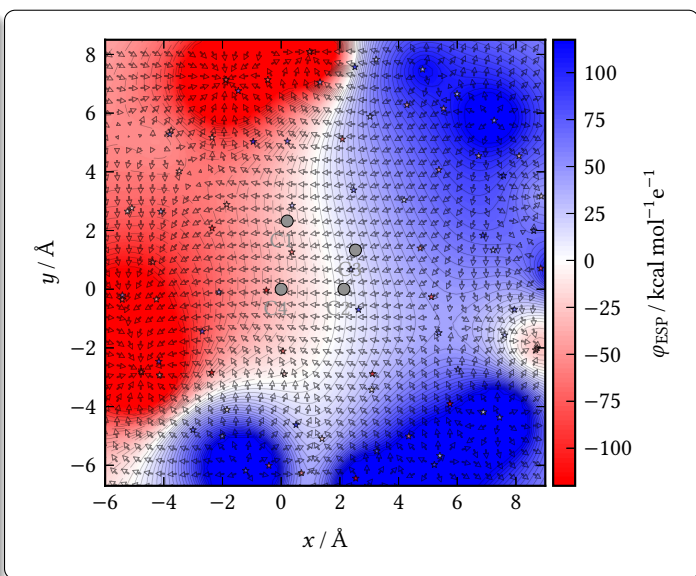


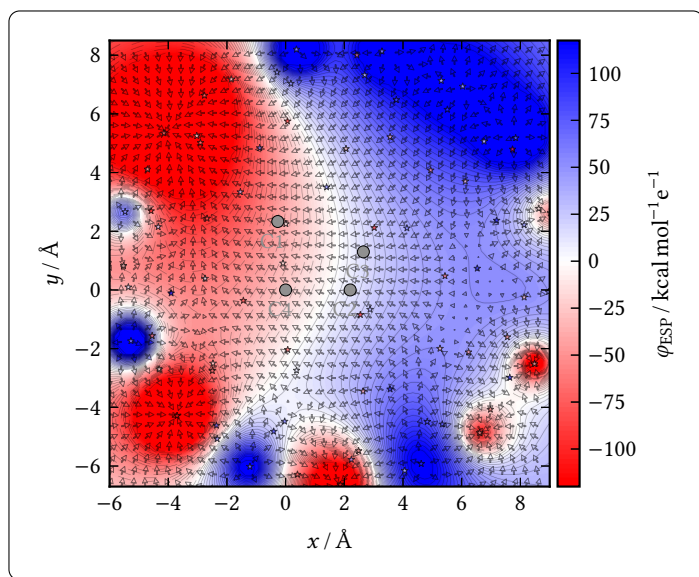
Fig. 7.13: $N_{\text{Ch}} = 81$ (sphere, adaptive): GOCAT r1 of Fig. 7.10 stabilizing a strongly catalyzed *endo* DA reaction showing an asynchronous two-step mechanism *via* a zwitterionic intermediate. The plotting details are the same as in Fig. 7.11.



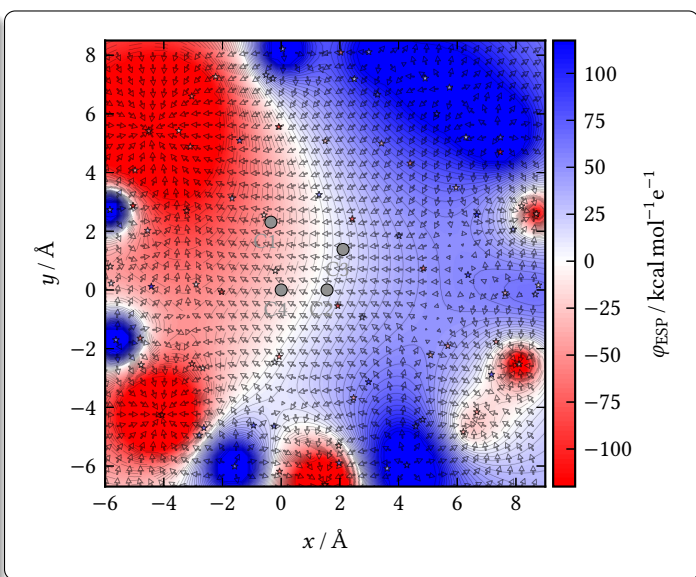
(a) r987, TS frame (cf. Fig. 7.11(b))



(b) r6, TS frame (cf. Fig. 7.12(b))



(c) r1, 1st TS frame (cf. Fig. 7.13(b))



(d) r1, 2nd TS frame (cf. Fig. 7.13(c))

Fig. 7.14: $N_{\text{Ch}} = 81$ (sphere, adaptive): ESP, φ_{ESP} , and its EF as the negative gradient plotted as arrows, projected onto the plane of the four atoms (C1–4) for the TS frames of the *endo* DA reaction shown in Figs. 7.11 to 7.13. For plotting details, see Fig. 7.9.

7.4.3 Uniform Electric Fields and Comparison

Uniform field effects, which are already well investigated for these DA reactions, including other derivatives and reactions (see Section 7.1), can serve as a baseline approach in order to discriminate these uniform fields from the globally optimized and non-uniform ones of the GOCATs. Moreover, an effective mechanism selection enforced by the global optimization and fitness definition in contrast to the simple SP trends in the literature without optimizations can also be discussed on this basis more easily. Therefore, reaction paths of uniform z -fields are shown in Fig. 7.15. These are the mentioned “plate capacitor” GOCATs creating a uniform field in, *e.g.*, the z -direction. One such plate capacitor GOCAT is plotted in Fig. A.18 on p. 282 in the Appendix. Furthermore, complementary to the ones discussed in this Section, more electric field plots along the other Cartesian axes are given in Appendix A.5 on pp. 282ff.

Both field directions were generated and can clearly demonstrate the trend of absolutely oriented uniform electric fields onto the energies and gradients. The reaction barrier decreases a lot until a value of about $\Delta\varphi_{\text{ESP-TS-}z\text{-pl}} \approx 69.0 \text{ kcal mol}^{-1} \text{ e}^{-1}$ (creating fields of $F_{\text{EF-TS-}z\text{-pl}} \approx -17.3 \text{ kcal mol}^{-1} \text{ \AA}^{-1} \text{ e}^{-1} = -74.9 \text{ MV cm}^{-1} = -1.46 \cdot 10^{-3} \text{ au}$) where the reaction barrier between the former R and TS frames disappears completely. Then, even

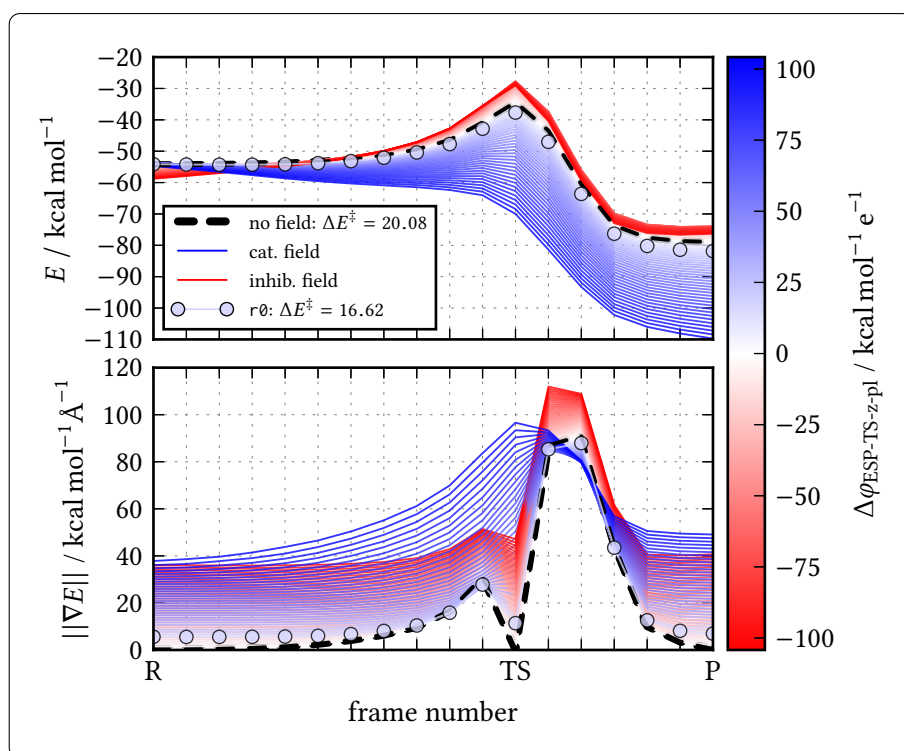


Fig. 7.15: Plate GOCATs (static): Reaction energy profiles after applying a uniform electric field in z -direction only (no other fields present) with incrementally increasing field strengths in both directions for catalysis and inhibition. Remark: The reaction profiles in small field strengths, *i.e.*, small $\Delta\varphi_{\text{ESP}}$ around $0 \text{ kcal mol}^{-1} \text{ e}^{-1}$, are (almost) white. Also the strength of the EF for the uniform $r\theta$ is only small (light blue) and its energies similar to the reference. Barriers, ΔE^\ddagger , are given in kcal mol^{-1} .

more extreme fields follow that are stronger than the ones found in any optimized **GOCAT** above.

Additionally, and more importantly with respect to the optimized **GOCATs**, we can see another simple trend: By increase of the field in the z -direction, the gradient norms substantially increase, especially at the **TS** frame. This is as expected because the pristine gas phase **TS** must become unstable such that the actual **TS** relaxed on the effective **PES** in the field shows the asymmetry and the two-step mechanism, finally. If the frames are not allowed to relax—as in Fig. 7.15—the best uniform plate **GOCAT**, $r\theta$, separately annotated in the Figure, shows a barrier of $\Delta E^\ddagger = 16.62 \text{ kcal mol}^{-1}$ and a substantially lower field, which is completely in compliance with, *e.g.*, the spherical static **GOCAT** $r\theta$. Thus, this uniform $r\theta$ can be seen as a limiting case of the maximal affordable strengths of uniform fields in a static fitness function without relaxation *via* **NEB** for this **DA** reaction. Every stronger field (and naturally all inhibiting fields since inhibition results in huge fitness values as this was not the optimization objective) is penalized heavily due to the gradient norm increase, the slight **R** destabilization and even some shifts of the minimum energy frame before the **TS**. That is, also the discrete penalty for the stationary point shifts already sets in heavily in most uniform fields.

The differences between this $r\theta$ (uniform) and $r\theta$ in the spheres (static) can then be ascribed to both non- z and non-uniform or local effects. This is even more instantiated in the **vdW** case, where separate atoms can be addressed by the charges, as discussed above. Note also that a uniform z -field indeed already destabilizes **R** slightly, *i.e.*, $\Delta E_{\text{R}} \in [0, 0.63] \text{ kcal mol}^{-1}$ for $0 \leq \Delta\varphi_{\text{ESP-TS-z-pl}} \leq 100 \text{ kcal mol}^{-1} \text{e}^{-1}$. Hence, **GA**-optimized **GOCATs** are highly motivated to evolve *some* field components in other directions than z and/or inhomogeneity in order to reach $\Delta E_{\text{R}} \leq 0 \text{ kcal mol}^{-1}$ again, which is indeed realized in Figs. 7.3 and 7.7; both y - and x -fields suffice here for a global path shift, see Fig. A.19 on p. 283. By using multiple handles and not just one uniform field, the optimized **GOCATs** show a lower barrier, *i.e.*, a better fitness, than the uniform plate **GOCATs**.

Finally, several properties are compiled in Table 7.1 on the next page. The electrostatic properties of which many were already discussed are given in the first block. Furthermore, the Table also includes other field directions and the barrier estimation by a simple polarized dipole approximation by use of Eq. (6.2) on p. 171 in second block. This equation was used here again after projecting the fields as well as the polarized molecular dipole moments onto the normalized direction vectors indexed as “3m1-2m4”, where both projected fields were additionally averaged (*cf.* Eq. (7.5) on p. 184). Then, the stabilization energies, which are summarized in the third block of the Table, can be estimated and compared to the, in fact, occurring energies that are given in the fourth block.

The points to be stressed here are the following: The fields indeed increase a lot from “left to right” in the Table for the **GA GOCATs** (sphere, **vdW**, adaptives). Note that the plane descriptors (**TS-z-pl**) are less useful for **vdW** as the charges can by chance sit very nearby or even inside the bounding box. Therefore, “3m1-2m4” is more adequate in this case. Then, the simple trend of enlarging field sizes also holds for this descriptor. The three ranks of the plate **GOCATs**, $r\theta$, $r25$ and $r65$, were selected to show a similar barrier

Table 7.1: Some properties of selected **GOCATs** at their **TS** frames. All energies are given in kcal mol⁻¹, gradients in kcal mol⁻¹Å⁻¹, **ESP** values in kcal mol⁻¹e⁻¹, projected **EF** in kcal mol⁻¹Å⁻¹e⁻¹ and projected dipole values in D. In the order of the columns, the **GOCATs** are: r0 (sphere: Fig. 7.4), r0 (vdW: Fig. 7.8), the adaptive solutions (Figs. 7.11 to 7.13, the 1st **TS** for r1) and three selected uniform field **GOCATs** from Fig. 7.15, including the plate **GOCAT** r0.

property	selected GOCAT							
	r0 (sphere)	r0 (vdW)	r987 (sphere, adaptive)	r6	r1	r0 (uniform z-field)	r27	r65
ΔE_{\ddagger}^a	16.23	13.32	11.00	5.56	3.40	16.62	11.63	3.12
f^b	180.52	149.54	123.14	61.43	54.84	192.84	2869.12	9831.66
average $\ \nabla E_{\{R, TS, P\}}\ ^c$	8.53	10.56	2.26	1.37	1.82	7.98	19.11	38.25
$\Delta\phi_{\text{ESP-TS-z-pl}}^d$	15.17	9.13	50.43	70.22	86.79	16.10	36.68	69.00
$\Delta\phi_{\text{ESP-TS-y-pl}}^d$	-5.39	-15.36	-2.71	30.10	45.90	0.00	0.00	0.00
$\Delta\phi_{\text{ESP-TS-x-pl}}^d$	-1.09	5.67	5.94	22.60	-16.96	0.00	0.00	0.00
$F_{\text{EF-TS-z-pl}}^e$	-3.80	-2.29	-12.28	-16.31	-18.46	-4.03	-9.18	-17.27
$\Delta\phi_{\text{ESP-TS-3m1-2m4}}^d$	6.71	8.60	25.20	35.39	47.23	8.44	19.22	36.16
$F_{\text{EF-TS-3m1-2m4}}^e$	-3.13	-4.01	-11.45	-15.21	-17.53	-3.94	-8.97	-16.88
$\mu_{\text{R-3m1}}^f$	0.04	0.16	-2.91	-5.60	-7.67	0.00	-0.10	-0.30
$F_{\text{EF-R-3m1}}^e$	-2.85	-1.17	-11.03	-16.42	-21.11	-4.02	-9.15	-17.21
$\Delta E_{\text{R-3m1}}^g$	0.02	0.04	-6.68	-19.14	-33.70	0.00	-0.19	-1.08
$\mu_{\text{TS-3m1}}^f$	-4.12	-4.06	-5.80	-7.40	-8.80	-4.11	-5.03	-6.62
$F_{\text{EF-TS-3m1}}^e$	-3.10	-3.52	-10.99	-14.58	-19.35	-3.94	-8.97	-16.88
$\Delta E_{\text{TS-3m1}}^g$	-2.66	-2.98	-13.27	-22.47	-35.46	-3.37	-9.40	-23.26
$\mu_{\text{R-2m4}}^f$	0.04	0.11	-3.55	-6.43	-7.29	0.00	-0.10	-0.30
$F_{\text{EF-R-2m4}}^e$	-3.37	-2.14	-11.32	-16.31	-18.95	-4.02	-9.15	-17.21
$\Delta E_{\text{R-2m4}}^g$	0.03	0.05	-8.37	-21.83	-28.76	0.00	-0.19	-1.08
$\mu_{\text{TS-2m4}}^f$	-4.13	-4.12	-6.01	-8.51	-8.43	-4.11	-5.03	-6.62
$F_{\text{EF-TS-2m4}}^e$	-3.17	-4.51	-11.91	-15.85	-15.71	-3.94	-8.97	-16.88
$\Delta E_{\text{TS-2m4}}^g$	-2.72	-3.87	-14.90	-28.08	-27.57	-3.37	-9.41	-23.28
$\Delta E_{\text{R-3m1-2m4}}^g$	0.03	0.05	-7.53	-20.48	-31.23	0.00	-0.19	-1.08
$\Delta E_{\text{TS-3m1-2m4}}^g$	-2.69	-3.42	-14.09	-25.27	-31.51	-3.37	-9.40	-23.27
$\Delta\Delta E_{\text{3m1-2m4}}^{\ddagger h}$	-2.72	-3.47	-6.56	-4.79	-0.28	-3.37	-9.21	-22.18
ΔE_{R}^i	-0.02	-0.01	-3.75	-26.25	-35.06	0.31	0.56	0.59
ΔE_{TS}^i	-4.04	-6.90	-12.83	-40.93	-51.74	-3.31	-8.63	-19.62
$\Delta\Delta E_{\text{static}}^{\ddagger a}$	-4.01	-6.90	-9.08	-14.68	-16.68	-3.62	-9.18	-20.21

- ^a ΔE_{\ddagger} is the final (effective, if a small shift of the lowest energy frame occurs) energy barrier as already shown in all other contexts; $\Delta\Delta E_{\text{static}}^{\ddagger} = \Delta E_{\text{static}}^{\ddagger} - \Delta E_{\text{ref}}^{\ddagger} = \Delta E_{\text{TS}} - \Delta E_{\text{R}}$ is the barrier decrease always between the first frame, **R**, and the **TS** frame such that $\Delta E_{\ddagger} \geq \Delta E_{\text{static}}^{\ddagger}$.
- ^b Fitness of the respective **GOCAT**; the pristine gas phase path without any **GOCAT** has $f_{\text{ref}} = 220.87$.
- ^c Mean gradient norm of all three stationary frames (**R**, **TS**, **P**).
- ^d Difference in **ESP** along the *z*-, *y*- or *x*-plane directions, cf. Figs. 7.1 and 7.2. A positive difference, *i.e.*, a positive voltage, is the amount of work per charge needed to move a positive charge “uphill” in **ESP** against a field pointing “downwards”, *e.g.*, in the negative *z*-axis direction.
- ^e The corresponding *mean EF* projected onto the indexed direction. For instance, for the *z*-plane, each 2-point entity of the $\phi_{\text{ESP-TS-z-pl}}$ was simply divided by the distance and averaged for the total grid (cf. Fig. 7.2(e), Eq. (7.5), and see depictions in the main text).
- ^f Polarized molecular dipole moment within the **GOCAT** projected onto the given (normalized) direction as vector between the specified atoms (*e.g.*, “3m1”, meaning the direction vector: C3-atom minus C1-atom).
- ^g Estimated energy stabilization (negative) with the simple formula for the projected (or local) field/dipole moments $\Delta E = -\mu_{\text{proj}} \cdot F_{\text{proj}}$ for either the respective C-C atoms direction vector indexed or the averaged version (“3m1-2m4”) of both the former energy values.
- ^h Estimated energy barrier decrease computed *via* Eq. (6.2), using the projected scalar **EF** and bond dipole moments and averaging both symmetric “3m1” and “2m4” bond contributions.
- ⁱ In fact calculated energy stabilization at **R** or **TS** as $\Delta E_{\{R, TS\}} = E_{\{R, TS\}} - E_{\text{ref}, \{R, TS\}}$ relative to the reference gas phase energies.

to the spherical ones. Clearly, they do not show any field along the y - or x -direction, by construction, conversely to all other optimized **GOCATs**. However, note the similarity of the uniform $r\theta$ and the spherical $r\theta$ with regard to most of the properties in the first upper block, e.g., f , $\|\nabla E_{\{R,TS,P\}}\|$, $F_{EF-TS-z-pl}$; the non- z -field and the other smaller inhomogeneity enhance the spherical **GOCAT** slightly compared to the plate **GOCAT**. The inhomogeneous effects can furthermore be traced back to the change in properties from the spherical $r\theta$ to **vdW** case. The latter is even better, but by using more of the local and non- z impacts.

Looking at the estimated $\Delta\Delta E_{3m1-2m4}^\ddagger$ (Eq. (6.2)) and in fact calculated $\Delta\Delta E_{static}^\ddagger$ (third to fourth block), one can discern the uniform **EF** from the rest. As the trend and estimation is quite good in all the plate **GOCATs** ($\Delta\Delta E_{3m1-2m4}^\ddagger \in [-3.37, -23.27]$ kcal mol⁻¹ vs. $\Delta\Delta E_{static}^\ddagger \in [-3.62, -20.21]$ kcal mol⁻¹), there are some bigger differences in the spherical and **vdW** cases. This is due to the fact that when having fields in other directions than z (which is almost parallel to both “3m1” and “4m2”), these components will not be present in the estimated barrier decrease. Moreover, if fields are locally *changing* (inhomogeneous), these are averaged away and, consequently, this might also lead to the differences. Inhomogeneous fields would necessitate higher-order moments and derivatives of the field to be used for such computations.

However, such estimations for the uniform plate **GOCATs** are quite solid, at least for small field strengths, as shown in Fig. 7.16 on the following page, using both the *polarized* molecular dipole moment as well as the permanent ones from the gas phase path before. That is, in the latter case, just by looking at the molecular dipole moments *a priori*, one can already detect a small projection onto the z -axis (or C–C-bonds) for **R** and a large one for **TS**, and thus one can calculate an energy barrier decrease by using the simple dipole energy expression of Eq. (6.2). For the optimized **GOCATs**, such a trend is also clearly detectable but less distinct due to the higher complexity of the fields; this is shown in Fig. 7.17 on the next page.¹⁶

Looking now at the properties of the adaptive **GOCATs** in Table 7.1, one can see, despite the appearance of more straight z -fields which were discussed and shown above (Figs. 7.11 to 7.13), also high **ESP** differences in other directions.¹⁷ With relaxation of the structures, we see a severe **R** stabilization that is not present in any static case. Often, this merely results from a small rotation of the **R** frame in the field causing a higher dipole coupling for the **R** frame, *i.e.*, the dipole is re-oriented in the external electric field for optimal stabilization, as expected. Accordingly, the discussed longer reaction coordinate results.

Furthermore, the estimation with Eq. (6.2) fails almost completely in these strong, non-uniform fields that induce high polarizations of the structures and different **MEPs**, in some but not in the majority of cases (see Section 7.5). The reasons could be manifold. Electron density reorganization synchronously to the mechanistic change (zwitterionic intermediate) could happen besides the main one along the “3m1” and “2m4” bonds, which

¹⁶Note that some estimations even would lead to a barrier *increase*, *i.e.*, to inhibition (abscissa) in contradiction to the in fact observed barrier decrease (ordinate) in Fig. 7.17.

¹⁷The properties were calculated for each *different* structure in each **GOCAT**. Thus, each bounding box (for the plane descriptors) will vary and thus also the distance between those, which results in different conversions to final **EFs**.

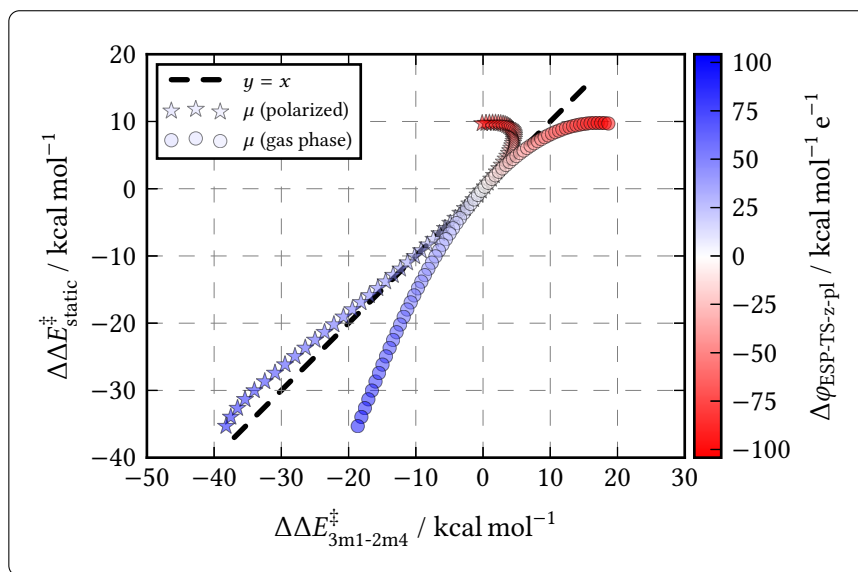


Fig. 7.16: Plate GOCATs (static): In fact calculated barrier decrease, $\Delta\Delta E_{\text{static}}^{\ddagger}$ (always between the R and TS frames of the pristine gas phase path), vs. the estimated one from Eq. (6.2), $\Delta\Delta E_{3m1-2m4}^{\ddagger}$, by using the average of the projected EF, F_{EF} , and molecular dipole moments, μ , onto the direction of the C–C-bond creation (“3m1-2m4”, cf. Fig. 7.2(f)). Here, either the fully EF-polarized, projected dipole moment $\mu_{\text{TS,proj}}$ or the unpolarized one from the pristine gas phase was used. The dashed line illustrates perfect association.

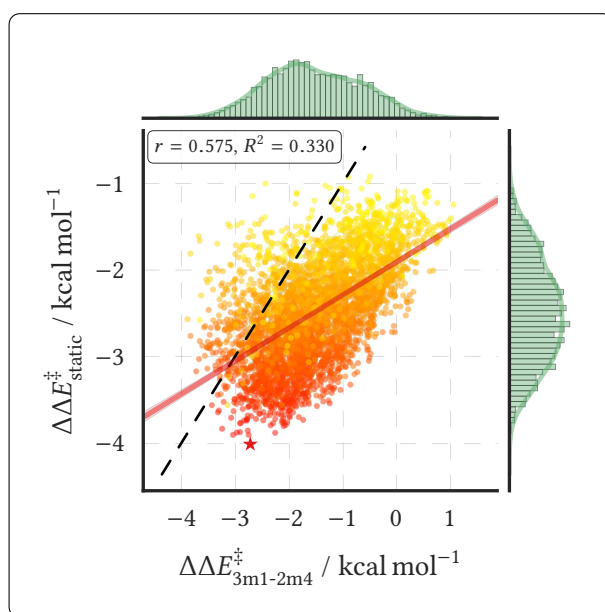


Fig. 7.17: $N_{\text{Ch}} = 81$ (sphere, static): Correlation plot of the barrier decrease vs. the estimated one by Eq. (6.2), compare with Fig. 7.16. The complete population is shown with $N_{\text{GOCAT}} = 4070$ individuals; $r\theta$ is emphasized by a red star. The black dashed line shows a perfect linear trend without an intercept, whereas the red line corresponds to the linear regression of the data.

deteriorates the significance of the descriptors, *i.e.*, the local projections. Furthermore, with any inhomogeneous field, higher-order moments in addition to the dipoles would be needed to completely assess the energetic influence by the electric field (Section 2.3.2). Besides, the dipole-based estimation also worsens its predictive power or even fails in very strong fields for the uniform plate **GOCATs** (*cf.* Fig. 7.16).¹⁸

7.5 Discussion

7.5.1 Comparison with Literature

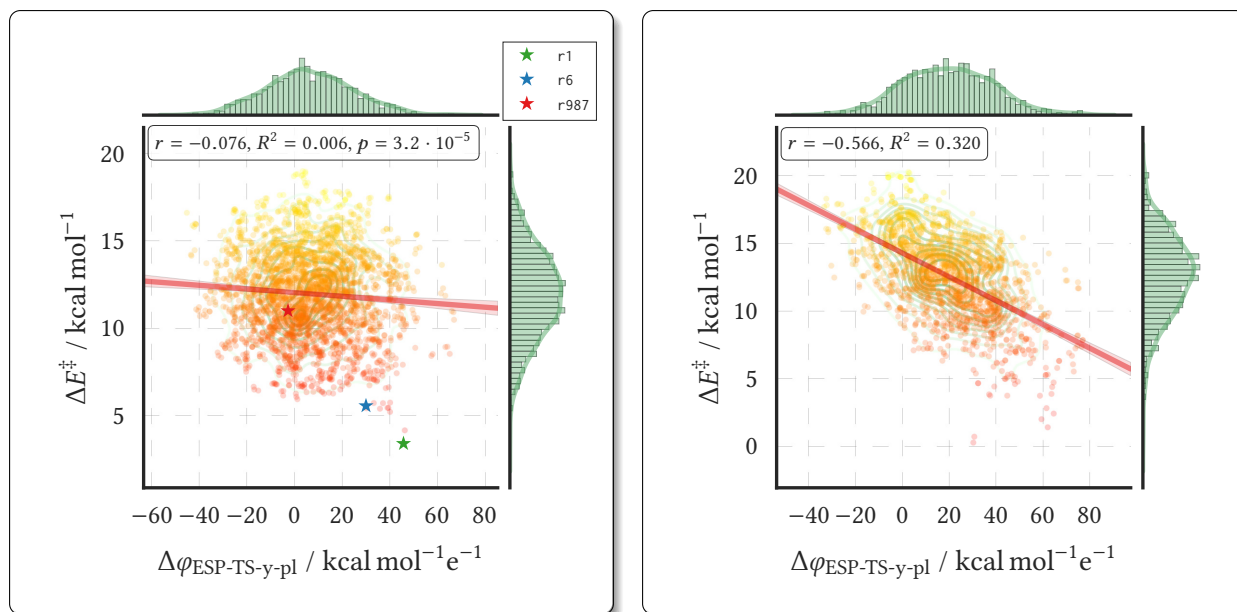
As introduced in Sections 7.1 and 7.2, an **OEEF** with $F_z < 0$ was reported by MEIR *et al.*^[459,463]¹⁹ to lower the barrier up to $\Delta\Delta E^\ddagger \approx -8 \text{ kcal mol}^{-1}$ at $F_z = -0.0125 \text{ au}$, while $F_z > 0$ inhibits the reaction by about $6.4 \text{ kcal mol}^{-1}$ (which was not the optimization target in the present study, but which is also included in Fig. 7.15). Looking at the plate **GOCATs**, in the present case on **PM7**, we see a barrier decrease of $\Delta\Delta E_{\text{static}}^\ddagger = -16.91 \text{ kcal mol}^{-1}$ in a field of $F_{\text{EF-TS-z-pl}} = -15.06 \text{ kcal mol}^{-1} \text{ \AA}^{-1} = -0.0127 \text{ au}$. In **PM7**, the gas phase barriers are already at $\Delta E^\ddagger = 20.07 \text{ kcal mol}^{-1}$ (21.13 for *exo*), *i.e.*, a bit higher than on the **DFT** level with $\Delta E^\ddagger = 15.5 \text{ kcal mol}^{-1}$ (16.7).^[463] The results are thus in qualitative agreement, while the main difference stems from the **R** stabilization in the literature of $\Delta E_{\text{R}} = -11.0 \text{ kcal mol}^{-1}$ against the present *destabilization* of $\Delta E_{\text{R}} = 0.63 \text{ kcal mol}^{-1}$. The projection of the molecular dipole in **PM7** onto the *z*-axis is almost zero, $\mu_{\text{R-z}} = -0.24 \text{ D}$, whereas Ref. [463] found a more significant contribution of up to $\mu_{\text{R-z}} = -5.76 \text{ D}$ (the ones at the **TS** structure are similar and higher, of course, needed for the barrier decrease). As the dipole moment norms are more similar (**PM7**: 4.77/8.42 vs. **DFT**: 7.84/9.66 for **R/TS**), this field influence on **R** stems from the precise alignment of the fields and the superposed molecular frames. In all **GOCATs**, there is *one* surrounding for all frames. Probably, this is also true for Ref. [463] defining a “uniform reaction axis”, as they say. Thus, this difference shall be traced back to the exact angle between the diene and dienophile. In **PM7**, the **R** dipole lies almost orthogonal to the *z*-axis, which results in (almost) $\Delta E_{\text{R}} \approx 0.0$. Compared to **DFT**, one must thus state that the possibility of **R** stabilization vs. destabilization for the **DA** reaction is notably dependent on the level of theory and dependent on the specific **MEP** structures on **PM7**, including their molecular dipoles.²⁰

Moreover, the *x*-axis orthogonal to the symmetry plane of the pristine gas phase path does not show any barrier decrease, as all frames are simply vertically shifted;^[463] for the

¹⁸Emulating uniform fields by plates also generates some small inhomogeneous fields that scale with the overall field strengths; a strict dipole coupling of a field into the Hamiltonian available in other program packages would remedy this situation. Additionally, the calculated **PM7** dipole moments themselves are approximated (semi-empirically) and are based partially on the effective atomic charges and corrections.^[478] In this case, dipole moments are likely wrongly estimated in very highly polarizing electric fields.

¹⁹They also used **SPs** as well as specific **TS** optimizations on B3LYP/6-311++G(d,p)/BP86/6-31+G(d) level of theory, with fully uniform fields in specific directions of a similar Cartesian coordinate system as in Fig. 7.1, but with some other sign convention; hence, the signs from the literature will be changed, if needed, to be aligned with the ones of the present contribution.

²⁰Hence, also the $\Delta E_{\text{R}} > 0$ search space for the **DA** reaction, which is shown in Fig. A.13 in the Appendix, might be less important or even absent on other levels.



(a) *endo*: barrier decrease along y

(b) *exo*: barrier decrease along y

Fig. 7.18: $N_{\text{Ch}} = 81$ (sphere, adaptive): Correlations of $N_{\text{GOCAT}} = 2952$ adaptive GOCATs with mean $\|\nabla E\| < 5 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ along the y -axis for the *endo* case in Fig. (a) (*endo* was shown in *all* Figures so far). For comparison, correlations for $N_{\text{GOCAT}} = 1776$ adaptive GOCATs with mean $\|\nabla E\| < 5 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ along the y -axis for the *exo* diastereomer are given in Fig. (b).

plate GOCAT, this is recapitulated in Fig. A.19 in the Appendix. There is also no correlation trend at all for this axis in any of the static or adaptive, optimized GOCAT (not shown). For y , a different influence on *exo* vs. *endo* was reported,^[463] leading to a $\Delta E^\ddagger = 13.9$ vs. $15.3 \text{ kcal mol}^{-1}$ at the highest field of $F_y = -0.0125 \text{ au}$, respectively. Because of this differing EF influence, the *exo* case can even be faster than the usually kinetically dominating *endo* case; this is the demonstrated stereoselectivity by the OEEF.^[463] In the GOCATs, there is usually always a variance in the x - and y -direction, as already discussed for Table 7.1, due to the inhomogeneity and the mixed field components at the same time, including the noise (or subtle) effects. Yet, within the adaptive GOCATs with increasing field strengths and bigger possible influences on the reaction, the possible diastereoselectivity is also visible in Fig. 7.18. This Figure compares the y -influence on the *endo* vs. the *exo* case.

With the TS relaxation in the EF in Refs. [459, 463], they find $\Delta d_{\text{async}} = 0.337 \text{ \AA}$ for the strongest field, $F_z = -0.0125 \text{ au}$, showing still a concerted mechanism. Adding (implicit IEF-PCM)^[479] CH_2Cl_2 , the asynchronicity reaches $\Delta d_{\text{async}} = 0.887 \text{ \AA}$ for intermediate fields, $F_z = -0.0075 \text{ au} = -8.89 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, and finally the two-step mechanism in a field of $F_z = -0.0125 \text{ au} = -14.82 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ and with the help of the solvent.

In the adaptive GOCATs, in accordance, we find also similar asynchronous Δd_{async} values; these are illustrated in Fig. A.17 on p. 281 in the Appendix. The selected GOCATs shown in this Section are similar but in slightly stronger fields, *i.e.*, r_6 (one-step) in a field of $F_z = -16.31 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ and r_1 (two-step) in $F_z = -18.46 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ (*cf.* Table 7.1). However, the GOCAT model does not need an additional solvent because the full electrostatic embedding is flexible enough and adapts already to the needs of the

core structures. Some non-uniform fields are thus probably needed and introduced in the literature studies,^[459,463] essentially, to stabilize the zwitterionic intermediate.²¹ All these trends are thus completely in line and nicely reproduced by the (almost) bias-free GA optimization of the EFs around the structures, finally reaching strong fields that are, however, still in a meaningful and physiologically relevant region,^[122,127,268,462,463] without the further need of explicit/implicit solvents.²²

7.5.2 Background Statistics for the Adaptive Fitness Function

In order to get an impression of what is happening in the background for the adaptive setting, some summary statistics are given in the following.

Generally, the full relaxation of NEBs during the GA by using Algorithm 3.2 on p. 93 is even more computationally expensive (for example, compare with the descriptions of Section 6.3.2). In the case of the 6000 GOCATs that were finally optimized for this one setting (Section 7.4.2), of which there were three selected individuals shown, there are about several thousand SPs needed for *one* GOCAT for a fully converged NEB optimization if it is started. This only happens if the fitness is already in a catalytic region, with $f_{\text{GOCAT}} \leq f_{\text{ref}} = 220.87$ in this case. Hence, essentially the startup of each population and the progression is similar to the usual *static* case, but when reaching the threshold, some orders of magnitude more expensive fitness evaluations set in immediately. This means that the final runtime can even be less easily anticipated since the number of SP calculations is very dependent on the unique population evolution.

At the end, about $N_{\text{tot}} = 7.60 \cdot 10^6$ iterations could be executed, unevenly distributed over 10 separate GA pools, which maps to about $N_{\text{GOCAT}} = 2 \cdot N_{\text{tot}}$ of sampled candidate solutions (the initialization of the pool can be excluded in this assessment). However, only about $N_{\text{NEB}} = 2.59 \cdot 10^5$ (1.70% of N_{GOCAT}) total NEBs started during all these GA steps. With all the separate soft and hard thresholds defined in Algorithm 3.2, about $N_{\text{GOCAT, fine}} = 2.61 \cdot 10^4$ (0.172%) converged finally and met all the conditions; the sanity checks include, e.g., to treat evolved “kinks” of the path detectable by angles that are defined by Eq. (2.92) on p. 60 and are substantially different from linearity, i.e., $\cos \theta_i = 0$, or convergence issues in the EF in any other (sub-)NEB step. Therefore, most of the sampled N_{GOCAT} are still evaluated in a similar manner as in the static case, by using Algorithm 3.1, though, with some important modifications explained in the following.

Each converged GOCAT in a pool can subsequently be tweaked, for instance, also just slightly by a very tiny GOCAT mutation as exploitation. If this individual happens to lie in the catalytic region (or even show a smaller fitness than before), again a NEB relaxation is triggered. Otherwise, or if the NEB even fails to converge, the starting path that can be

²¹Note that with a frame-wise solvent, also an intermediate state on the MEP will fully be relaxed, which is not the case in *one* overall GOCAT embedding. That is, the zwitterionic intermediate can fulfill a conformational rotation out of the attack-plane.^[463]

²²One difference, also with regard to relaxations of the plate GOCATs discussed below, is the use of internal coordinates for the optimizations of the structures in Ref. [463]. Therefore, the field will properly be aligned during the whole optimization. In the GOCAT model with explicit point charges generating the field, we have external (absolute) coordinates (and an anchor point) and the structures can indeed re-orient in the field. This changes the projected fields and dipole components.

already different from the gas phase path might still compete in the complete pool if it is better than the worst current **GOCAT**. Thus, at all times, some gas phase path **GOCATs** that are the same as in the static version can compete with very well-converged ones of a successful relaxation, using Algorithm 3.2, and with some other non-converging ones. The paths of the last kind will not be returned (in a “broken form”) from the fitness function, but the fact of occurring problems during the **NEB** might signify some problematic impact of that particular **GOCAT**. In other words, when the full re-optimization is successful, the resulting **GOCAT** is very tightly optimized. When the optimization fails, however, the fitness function falls back to essentially the same as in the static case. Then, the resulting **GOCAT** usually shows slightly worse gradient norms, etc.

As a result, the final extreme fields are very probably an outcome of multiple such *consecutive* (small-step) **NEB** relaxation rounds since the starting gas phase path is still important—*i.e.*, the full path optimization is still reference-based—and the new paths are not reached in *one* **GOCAT** relaxation.²³ There were no benchmark calculations, yet, whether bigger direct **NEB** relaxations without hard pre-checks (thresholds) could lead to the same results. The preliminary tests until now showed that these consecutive treatments are important for reducing the computational costs since too extreme fields or moves with very big impacts result in infrequently converging paths. Also in adaptive **vdW** embeddings (not shown here), the very nearby sitting Coulomb “singularities” can wreak havoc on the optimizations such that the overall more homogeneous fields of spheres are needed. The spheres often are harder to interpret in the Cartesian domain and also show some “noise”. However, at the reaction center nearby the atoms themselves, the **EF** shows a smaller variance than in the **vdW** case, where nearby sitting charges, *i.e.*, sources and sinks of the **ESP**, produce the inhomogeneity.

7.5.3 Critical View and Improvements

Allowing the relaxation due to the adaptive **NEB** implementation indeed improves upon the often subtle trends of the frequently highly unique solutions seen in both static cases: **vdW** with very inhomogeneous fields but quite symmetric coordinates, and the spheres with more uniform ones but with superimposed noise and with a complete lack of interpretability in the Cartesian domain. With relaxations, the possible impact of any surrounding is amplified heavily, which leads to a better correlation again of the barrier and the fields in *z*-direction. Generally, the total reaction coordinate length between all frames increases as well as the asymmetry of both to be created CC-bonds with increasing field strength. Though, many solutions in the final population still show only a small change in the mechanism; that is, a big part of the population is still dominated by small structural changes in the fields and the asymmetric and two-step **GOCATs** are rather outliers that are found at the frontier (best

²³To prove such an assumption, full histories of the exact search space moves, **NEB** convergences and **EF** analyses would be needed; such histories are partially produced, also for **CSO** for instance, but it is still very cumbersome and would need further extensions in order to fully chase the actual appearance and progression of the intrinsic properties (*i.e.*, the genotypes) during the **GA**.

part) of the final solutions. These mentioned correlations are all shown in the Appendix in Fig. A.17 on p. 281.

With relaxation, the search space will enlarge a lot. Here, not just the GOCAT charges must find its perfect coordinates, but with each outer change, a whole new effective PES can be sampled and re-optimized upon subsequently. And each such non-vertical adaptation of the MEP can, in total, lead to big steps during the GA search of qualitatively other type than the strict ESP-based energy shifts by the surrounding in the static case. Hence, it can be assumed that even more structural changes, including, e.g., even stronger rotated and more loosely associated R frames for the DA reaction, could be observed with even more GA iterations or with further adaptations of the fitness function.

At the moment, due to the current implementation of the fitness function as used in this application, it is not allowed to reach paths with a higher R energy than the TS energy. Thus, “barrier-free” reactions or paths similar to the uniform plate GOCATs in very strong fields (Fig. 7.15) were simply not within the feasible set of solutions. Whether this is physically meaningful, should be considered again for the future problem at hand. Also the meta-parameters of the adaptive NEB optimization were set as a compromise between computational performance and final convergence.²⁴ Hence, sometimes small kinks in the path and relaxation issues especially at the R and P frames were often present. Thus, the settings should also be benchmarked (again) in the future. The acceptance or success rate of the mentioned 0.172% sane candidate solutions after NEB could then surely be increased.

In the benchmark set used for the adaptive NEB implementation (cf. Section 2.5.1.3),^[336] sometimes meta-parameters such as the resolution ratio ζ , the step-length, control parameters for the restarts needed (due to the non-conservative nature of the NEB forces) etc., either led to a smooth convergence or to a failure (as already discussed in that Section, too, Fig. 2.9). This is due to any issues in any of the included steps of the full adaptive NEB for fast TS search, the CI NEB or the “cosmetics” run for tightening the MEP used in Algorithm 3.2. Thus, it must simply be restated again that reaction path optimization is hard to be completely automated for each conceivable case. To correct the bias of the meta-parameters, even multiple different reaction path optimization or TS search routines could be used with optionally falling back between these routines in order to increase the robustness of the MEP search if one algorithm turns out to be unsuccessful. Interestingly, also the uniform plate GOCATs, as seen in Fig. 7.15 on p. 201, show convergence issues that are intensified in stronger fields but also occur (erratically) in weaker ones. Here, especially the R frame is problematic. In stronger fields, the anhydride becomes essentially orthogonal to the diene for optimal dipole alignment. For this big movement in harsh, i.e., strong field surroundings, very robust optimization routines are needed. Some more impressions of adaptive plate GOCATs in this regard are shown in Fig. A.21 on p. 285 in the Appendix.

For a further improvement of the success rate of these adaptive NEB GOCATs in the current NEB implementation, for example, one could simply use *different* settings of the NEB leading to multiple NEB convergence trials for *one* GOCAT in the fitness evaluation. Then,

²⁴One could definitely use “tighter” NEB settings that, however, would increase the computational expense a lot.

the converged one or the best of these could be returned. Instead, one could implement a separate **TS** optimization routine that works on a **TS** candidate of an antecedent **CI NEB** that might not have been fully converged before.

Also note that by choosing any meta-setting of the **NEB** that might turn out to be worse-performing for specific paths in strong electric fields do simply *not* survive during the **GA**.²⁵ In principle—assuming for the sake of this argument—that **PM7** could lead to systematic errors due to its approximations (*e.g.*, **QM/MM** coupling model, Section 2.4, and the minimal basis) such that specific mechanistic changes were simply excluded, this would essentially be “hidden”, since any final **GOCAT** is already based or selected on being well-performing on precisely this setting, including the level of theory and all other meta-parameters of the Algorithm 3.2. As a result, this whole methodology and the benchmarking of these settings should also include other reactions and levels of theory to investigate if any further *bias* is still present.

Besides, further improvements of this adaptive setting might primary focus on, *e.g.*, “path-aware” algorithms: For **DA** and using spheres, this was less a problem, but already in other examples, *e.g.*, in a Menshutkin reaction (not shown), **MEPs** were optimized that showed not only a subtle change of the mechanism but that also led to completely new reactions at all, *i.e.*, **R** and/or **P** changed entirely, which had nothing to do with the tackled chemical reaction. This is again one entrance of overfitting if such changes are not intended *via* another (future) objective function.²⁶ To this end, *configurations* should also be detected and part of the **GOCAT**’s intrinsic information, for instance. This could be implemented by using simple dissociation detections and additionally graph theory based connection schemes. One could check then if the optimized **GOCAT** still favors the same overall reaction type as the intended one or if it leads to something different. Additionally, with a “reaction path niching”, a competition of alike and different reaction mechanisms, *i.e.*, concerted until two-step ones, could be managed.²⁷ Regarding the **DA** example, maybe even different relational barriers of *exo* vs. *endo* could be controlled in the *same GA* population in this way.

Furthermore, the current adaptive **NEB** fitness function actually does *not* explicitly favor to find multiple steps. The **GOCAT** r1 showing a two-step mechanism is rather a result of a path with overall small gradient norms. Consequently, also the second **TS** is quite fine. However, it was *not* automatically relaxed with another **CI NEB** round. Notwithstanding, the overall **MEP** was still quite acceptable at the end, as shown in Figs. 7.10 and 7.13(c). Actually, an improved version of the fitness function of Algorithm 3.2 would need only a tiny adaptation—just reading out and maybe starting additional **CI**s at each intermediate local maximum in energy. However, further possible overfitting must then strictly be eliminated. Without additional sanity checks and “reaction path awareness” (*vide supra*), very huge

²⁵This means that the **GA**-optimized **GOCAT**s are superior to the uniform plate ones as only such candidate solutions will be present that *did not* have any of these problems due to the whole optimization itself. Maybe the central-symmetric potentials in spheres or the inhomogeneous effects are helpful here.

²⁶This was already observed by translating very fit **PM7 GOCAT**s to **DFT**-based ones with full relaxations, **TS** optimizations, etc., before the era of Algorithm 3.2.

²⁷This also extends to the analysis/evaluation scripts for the **GOCAT** optimization that cannot yet cluster “reaction-path awarely”.

mechanistic changes could be possible, in the best case, or maybe just meaningless **CI** optimizations at intermediate “kinks” that could be intensified erroneously in this way. This topic could be addressed by further future research. Already in Section 2.5.1.4, some additional possible improvements were mentioned that could make the optimization (even) more robust, such as the double-nudging,^[337] similar kink-control mechanisms^[300,338] and the mentioned specialized **NEB** optimization algorithms, besides others (*cf.* Section 2.5.1.4).

7.5.4 Conclusion

To conclude this Chapter, it was demonstrated that electric fields can very well catalyze this **DA** reaction that, according to the corner cases of electrostatic catalysis described in Section 6.3.3, shows both varying dipole moment directions and magnitudes such that a field must “anticipate” the **TS** electrostatics for optimal **TSS** and barrier decrease. In accordance with experimental^[472,475] and theoretical results^[268,459,463,465] from the literature, the main results of electrostatic catalysis in direction of the so-called “reaction axis”, the direction of bond-breaking and/or making, could also be quantitatively reproduced using both the static (vertical) **GOCAT** as well as the adaptive (non-vertical) **GOCAT** model with relaxations of the **MEP** within the external fields. This was used to investigate the next important step towards a less-restricted and better automated optimization of **GOCATs**.

One reoccurring issue were all the subtle (noise) effects due to the highly inhomogeneous **EFs** that are fundamentally different from what was used in the literature so far. The fact of having found these highly varying **GOCATs** during the global optimization in both Cartesian and **ESP** domains for the **DA** reaction is actually ambiguous: For a *translation* back to real molecular structures, this could be advantageous when having multiple different possibilities of realizations which stabilize different mechanisms by different field impacts. Conversely, for an interpretation and understanding of the actual effects, this is clearly disadvantageous with overly complex candidate solutions which are widespread on the fitness landscape. Indeed, also other non-local effects and fields in other directions than along the literature-known *z*-component, *i.e.*, along the reaction axis, were often present. Whether this maybe unnecessary complexity is advantageous in the future for exploring rather unexpected effects, needs further research. Otherwise, one could also enforce simplifications of the candidate solutions by adapted future fitness functions.

Nevertheless, the main catalytic effect and the fields along the *z*-direction were in fact highly correlated and these fields led to (extreme) catalytic effects resulting in essentially *no* barrier at all. With the decrease of the barrier, the **DA** reaction developed from following a concerted synchronous one-step mechanism to a two-step mechanism, since the new zwitterionic intermediate in the latter case is strongly stabilized by the appropriate **EF**. Also, a small *y*-field effect for a diastereoselectivity in line with the literature studies was found. Furthermore, the **NEB** relaxations overall led surprisingly well to the anticipated results without any further bias except for the discussed **GA** and meta-parameter settings as well as the starting gas phase path.

Note that some more general conclusions will be drawn in the next and final Chapter of this Thesis.

Conclusions

Before coming to the conclusions, some condensed further impressions about versatile applications of the **GOCAT** theme are illustrated next in Section 8.1. The main work is then summarized as well as concluded in Section 8.2, which encloses a general overall outline of the **GOCAT** project and an “executive summary” of the results. Finally, an outlook is contoured in Section 8.3 addressing further possible steps and applications of the framework.

8.1 Master’s Thesis of BEHRENS

In order to widen the perspectives of possible electrostatic **GOCATs**, more selected impressions stemming from the Master’s Thesis^[480] of BEHRENS are *very* briefly discussed. This Master’s Thesis took place under my supervision. Yet, of course, the aim of this Section is not to take credit for others’ achievements but to complement the conclusions of this Section. The *telling* title of this Thesis was

Global Optimization Of Abstract Catalysts: On The Road To Application.

So, the aim was to use mainly the same setting as in Chapter 6, *i.e.*, the *vertical* or *static* mode, *as is* and *extend* the application spectrum to more chemically notable reactions. In the end, these investigations included not only two enzyme systems, namely Kemp eliminase and ketosteroid isomerase, but also two steps of the Monsanto process involving a catalytic cycle around a rhodium transition-metal complex for acetic acid production. These systems were studied with **PM7**, **GFN-xTB**¹ and partially also with **DFT**. The goal was also to investigate whether the electrostatic **GOCAT** model reaches its limits or whether understandable catalytic effects can again be observed. In the following, just some selected examples are recapitulated focusing on **PM7** and **GFN-xTB** results with regard to the Kemp eliminase enzyme and one step in the Monsanto process, respectively. For the other system, the ketosteroid isomerase, the reader is referred to, *e.g.*, Refs. [121, 122]. This ketosteroid isomerase is a naturally occurring enzyme and is (also experimentally) well characterized

¹ At the time of the Master’s Thesis, the successor GFN2-xTB were not yet available (*cf.* Section 2.2.3).

for electrostatic catalysis. **GOCATs** without any protein or theozyme backbone were able to equally well catalyze this reaction, as shown by BEHRENS,^[480] in general accordance with the measured trends and local dipoles.^[121,122]

Kemp eliminase: First, some selected Kemp eliminase results are shown in Figs. 8.1 and 8.2 on the next page and on p. 216. The specialty about this system is that it stems from a complete *ne novo* design by RÖTHLISBERGER *et al.* of an enzyme catalyzing this reaction without a natural counterpart before.^[111] These authors were able to create multiple enzyme variants, which were also refined further with directed evolution (as experimental technique) by KHERSONSKY *et al.* afterwards.^[481–483] Later, these were computationally re-investigated by BHOWMICK *et al.*,^[123,126] and they found that the electrostatic surrounding by the scaffold around the theozyme was *not* optimal yet in the oldest variants of RÖTHLISBERGER *et al.*, but this was (partially) improved in the experimental re-designs of KHERSONSKY *et al.* Hence, BHOWMICK *et al.* proposed to *focus* also on proper electrostatics of the scaffold in earlier stages in *de novo* designs, which cannot easily be corrected afterwards by the directed evolution steps otherwise.^[123] Consequently in Ref. [126], they followed exactly this idea, *i.e.*, the design of an enzyme using electric fields as optimization guidance criterion for single-site mutation steps.

This leads over to the **GOCAT** theme again. When such an electrostatic **GOCAT** is designed for this reaction, it should lead to the *globally* best electrostatic embedding possible, which was applied here to the Kemp eliminase variant KE59 of Ref. [111]. In contrast to the very detailed and intensive analyses of the **GOCATs** for, *e.g.*, the Menshutkin and the **DA** reaction (*cf.* Chapters 6 and 7), the results are very concisely sketched in the following:²

- Fig. 8.1(a): The expectation is clearly set for a rough direction of an electric field catalyzing the H-abstraction step by the glutamate side chain, the catalytic base Glu231, followed by the concerted ring opening.
- Fig. 8.1(b): The “full theozyme” system consists of nine amino acids in total around the reacting moiety and these are all included in the reaction coordinate. Obviously as expected, this “full theozyme” case is already better performing than the raw reaction with just the *one* base, Glu231, which case is named “no theozyme”. The reaction profiles are shown here without any **GOCATs** yet.
- Fig. 8.1(c): A spherical **GOCAT** generates an **EF** in accordance with the expected direction that is sketched in (a).
- Fig. 8.1(d): Another spherical **GOCAT** having just 10 charges (but still representing an outer scaffold, *i.e.*, sampled on a sphere) shows also a very similar **EF** in line with the expectations of (a) and the case of (c).

² For answering questions such as how good the **GOCATs** are compared to the full enzymes and/or to the computationally re-optimized ones by HEAD-GORDON *et al.*^[123,126] further studies and analyses are needed.

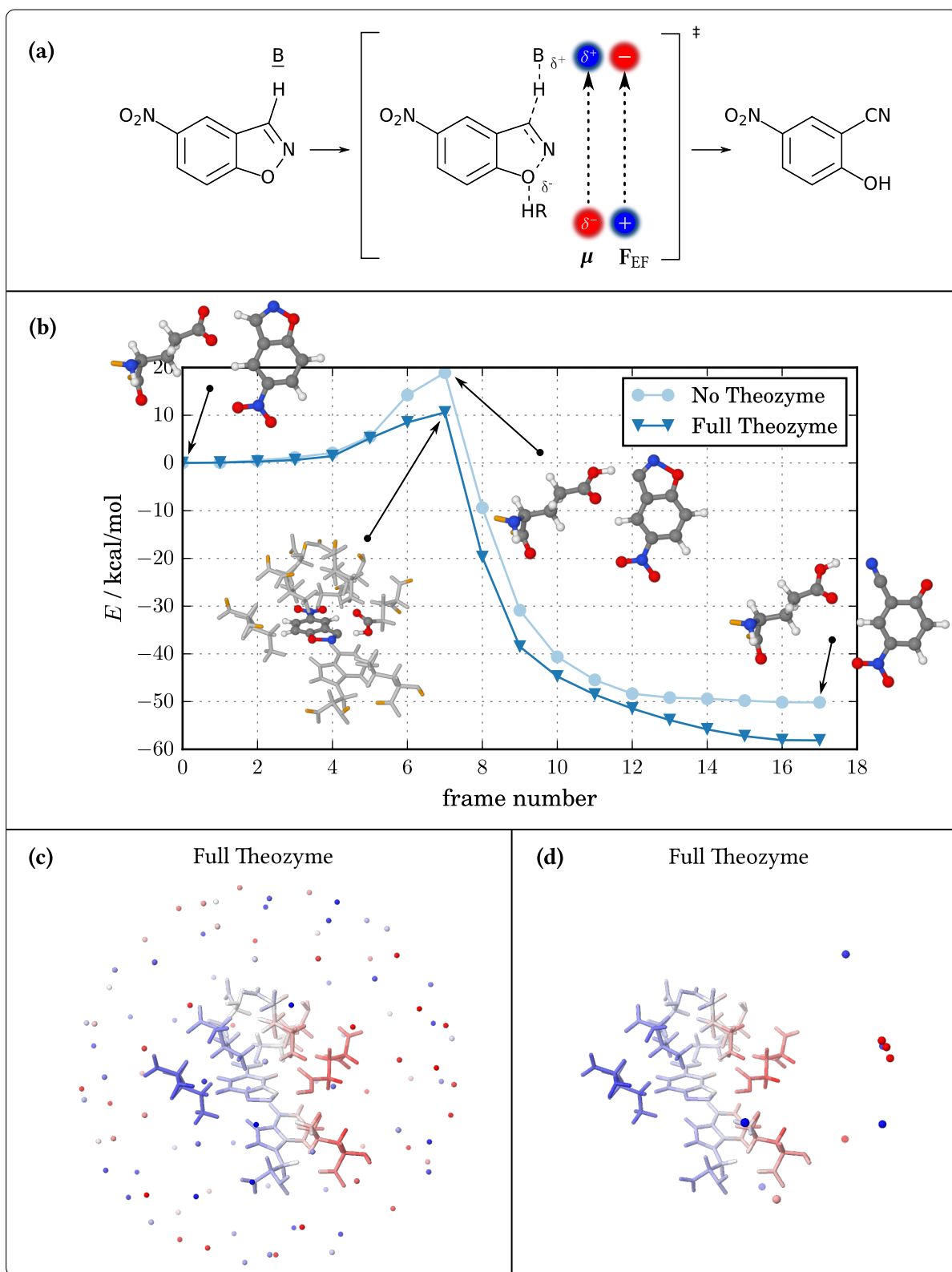


Fig. 8.1: PM7: Adapted data from Ref. [480], part I: (a) Kemp elimination with expected EF; (b) reaction energy profiles without a GOCAT; (c) “full theozyme” with a spherical uniform GOCAT, [red, blue] for $q_i \in [-0.3, 0.3] e$ and $\varphi_{ESP} \in [-12, 12] \text{ kcal mol}^{-1} e^{-1}$; (d) “full theozyme” with a spherical small GOCAT, [red, blue] for $q_i \in [-1, 1] e$ and $\varphi_{ESP} \in [-13, 13] \text{ kcal mol}^{-1} e^{-1}$. See the main text for explanations.

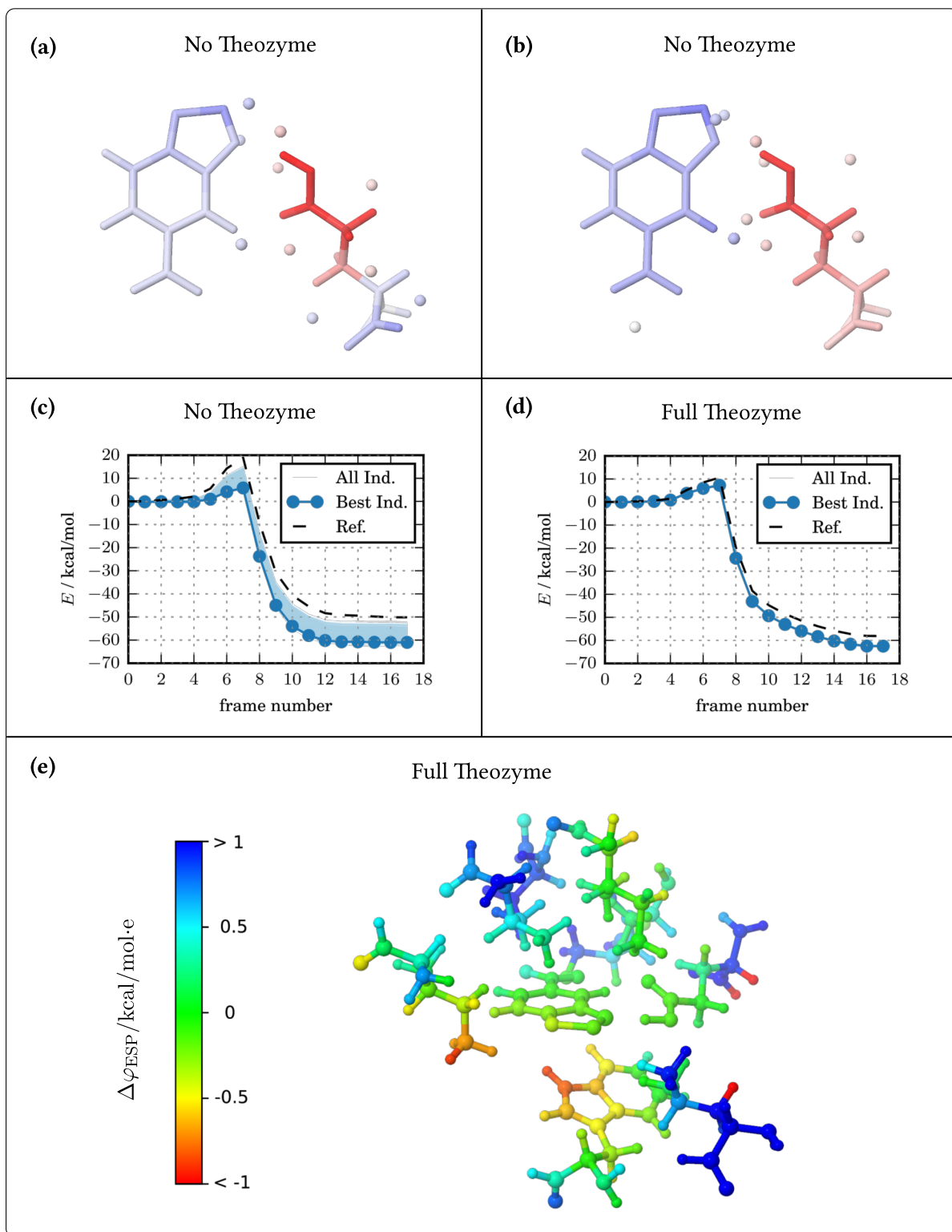


Fig. 8.2: PM7: Adapted data from Ref. [480], part II: (a) “no theozyme” with best GOCAT on vdW surface, [red, blue] for $q_i \in [-0.5, 0.5] e$ and $\varphi_{\text{ESP}} \in [-16, 16] \text{ kcal mol}^{-1} e^{-1}$; (b) “no theozyme” with another GOCAT colored as in (a); (c) energy profiles for all GOCATs, including (a) and (b); (d) energy profiles for all GOCATs, including Fig. 8.1(c); (e) mean ESP difference between all GOCATs in the “full theozyme” mode of Fig. 8.1(c)–(d). See the main text for explanations.

- Fig. 8.2(a): Without a full theozyme and optimizing a **GOCAT** directly on a **vdW** surface, the **EF** is still roughly oriented along the proposed dipole direction of the reaction scheme in Fig. 8.1(a). However, some positive (blue) **ESP** at the glutamate can be observed here.
- Fig. 8.2(b): This example is similar to the last one but does not show such a distinct local positive (blue) **ESP**, *i.e.*, it is more uniform.
- Fig. 8.2(c): Without a theozyme, the catalytic effect of a **GOCAT** can be greater, and this is shown here in comparison to the “full theozyme”. A barrier decrease of about $\Delta\Delta E^\ddagger \approx 13 \text{ kcal mol}^{-1}$ can be observed. The final barrier is even slightly smaller than the one of the best **GOCAT** in the “full theozyme” setting of Fig. 8.2(d).
- Fig. 8.2(d): This “full theozyme” mode does show a small but distinct barrier decrease of $\Delta\Delta E^\ddagger \approx 3 \text{ kcal mol}^{-1}$ by the spherical scaffold **GOCAT**. Note that in all these reaction energy profiles, the energies are shifted to 0 kcal mol^{-1} at the **R** frame. Thus, the local positive (blue) **ESP** feature of Fig. 8.2(a) leads to an increased **R**-destabilization as catalytic effect (which is, however, not visible here due to the shift). (As a remark: the spread in this Fig. 8.2(d) can simply not be seen since a loose niching has lead to a full convergence of the complete population to essentially one solution; the whole population that consists of multiple thousands of candidate solutions is plotted in all cases.)
- Fig. 8.2(e): Although the solutions in Cartesian space of Fig. 8.1(c)–(d) were obviously different, the resulting **EF** is not, which is accentuated in this Fig. 8.2(e) as a difference plot. This leads again to one simple picture in unison how to catalyze this reaction.

Monsanto process: As a second recapitulated investigation, the mentioned Monsanto process is shown in Fig. 8.3 on the next page (*e.g.*, compare with Refs. [484–487] for further computational studies on this system). It has to be stated as a caveat that such transition-metal complexes need even more thorough test calculations beforehand to validate the **SQC** levels of theory for the global optimization. Indeed, not only this Monsanto process but more transition-metal catalytic processes have been tackled first in the Master’s Thesis, but many systems showed artifacts of the **SQC** treatment compared to **DFT** calculations regarding already the pure **MEPs** without a **GOCAT**. Yet, the one step of the process shown here could rather be well described by the **GFN-xTB** method.

The first catalytic step, which is not shown here, is the oxidative addition of ICH_3 to the Rh-complex, which is also the rate-determining step. In fact, the final **GOCATs** that were optimized for this step in the Thesis^[480] were very similar to the ones concerning the Menshutkin $\text{S}_{\text{N}}2$ reaction (*cf.* Chapter 6) in many respects, including the resulting **ESP**. Comparing the Monsanto step with the Menshutkin reaction, ICH_3 takes over the role of ClCH_3 and $[\text{Rh}(\text{CO})_2\text{I}_2]^-$ the role of NH_3 . Strong catalytic effects with a barrier decrease of about $\Delta\Delta E^\ddagger \approx 11 \text{ kcal mol}^{-1}$ could also be observed for this step then. This is completely in line with standard organometallic textbook knowledge that compares such

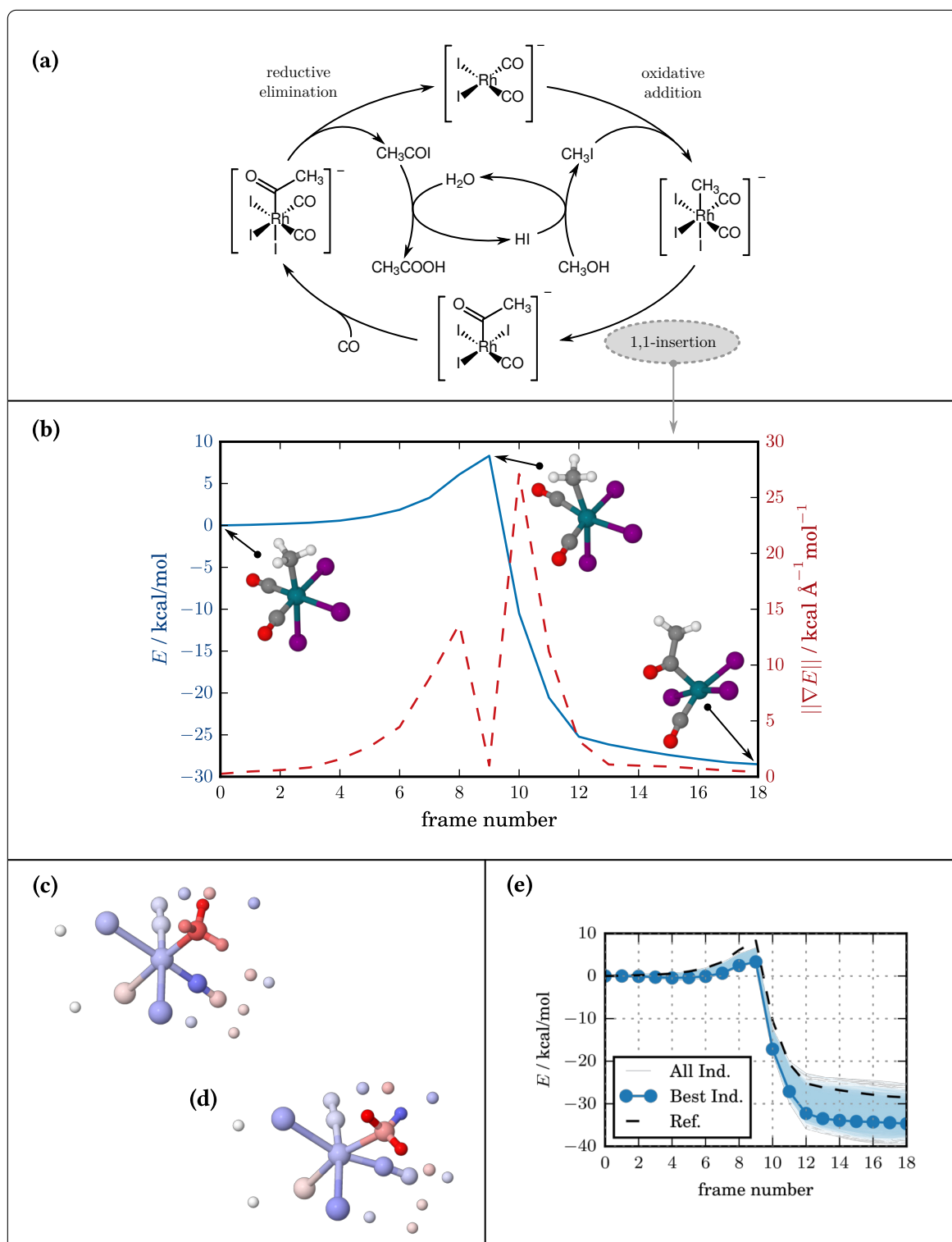


Fig. 8.3: GFN-xTB: Adapted data from Ref. [480], part III: (a) Monsanto process scheme; (b) reaction energy profile without a GOCAT for the insertion step; (c) R frame with the best GOCAT, [red, blue] for $q_i \in [-3, 3]$ e and $\varphi_{\text{ESP}} \in [-36, 36]$ kcal mol⁻¹ e⁻¹; (d) TS frame with the best GOCAT, coloring as in (c); (e) energy profiles for all GOCATs including the best individual. See the main text for explanations.

oxidative additions with S_N2 -like nucleophilic attacks on methyl halides (and others, see, e.g., Ref. [488, pp. 165f.]).

Therefore, the perhaps more interesting second insertion step is shown in Fig. 8.3; it can be summarized as follows:

- Fig. 8.3(a): The complete Monsanto process is sketched, of which the 1,1-insertion is further investigated in this Figure.
- Fig. 8.3(b): The reaction energy profiles show the small overall barrier of this insertion without a GOCAT.
- Fig. 8.3(c)-(d): The best GOCAT around the R and TS frames shows a negative (red) ESP at the methyl group and a positive (blue) one at the corresponding CO group as most significant feature, besides the positive (blue) ESP at the metal Rh center. It is known that such insertions become faster with increased electrophilicity of the metal center leading to a higher positive polarization at the C atom of the CO ligand. The overall step can be thought of as a migration of the (nucleophile) Me^- onto the carbonyl ligand (compare with, e.g., Ref. [488, pp. 187ff.]). This trend is clearly increased by the generated ESP. Though, also negative (red) values at other places are observed such that a description of stabilization and destabilization effects are less clear in this representation. Electron density is indeed shifted to the place where the new C–C bond between CO and CH_3 is formed, which was also investigated by BEHRENS.^[480]
- Fig. 8.3(e): The reaction energy profiles within the GOCATs that also include the best individual from (c)-(d) show a clear barrier decrease of about $\Delta\Delta E^\ddagger \approx 5 \text{ kcal mol}^{-1}$. As indicated, no simple single dipole direction can be assessed to this reaction that shows more subtle ligand-wise effects. In fact, the overall reaction energy profiles cover many solutions with both stabilization and destabilization effects at the R frame (which is not shown here due to the shift of all profiles to 0 kcal mol^{-1} at R again). And especially with positive (blue) ESP influences at the iodine ligands, strong stabilizations (at all frames) can be reached. More detailed discussions would need further research on this transition-metal reaction and are intentionally not pursued here.

8.2 Summary and Conclusion

A general outline and background of the approach taken in this Thesis is given in Section 8.2.1 first, and subsequently the most important findings of the investigations are summarized in Section 8.2.2.

8.2.1 General Project Outline

This Thesis dealt with abstract catalysis optimizations. One general problem occurring in all different flavors in chemistry is the *inverse* design of suitable chemical systems for target properties. Since the respective differential equation, the **time-independent Schrödinger equation (TISE)**, for any real-life problems cannot simply be inverted in a mathematical sense, usually proficient search algorithms and further representation methods of the chemistry are introduced to guide through chemical space of possible molecules. Thus, the inverse problem of designing chemical systems is often reformulated as an optimization problem. There are already some methods available to approach diverse property designs that are based on, for instance, **linear combination of atomic potentials (LCAP)**, alchemical potentials, diverse **machine learning (ML)** methods and different optimization schemes, including certain objectives for the catalytic optimization as in the **gradient-driven molecule construction (GDMC)** approach. Furthermore, some concepts for catalysis design have been developed, such as the **optimal catalytic fields (OCFs)** and the *theozymes* in the realm of enzymatic catalysis. These were all discussed in the Introduction in Chapter 1.

Due to the sheer extent of chemical space, fully deterministic enumeration of all feasible chemical molecules is strictly not possible, and a pure random search is evidently also not the most efficient approach. Capable metaheuristic global optimization algorithms were therefore leveraged in this Thesis, namely **evolutionary algorithms (EAs)**. In fact, in our work group we have gained a lot of experience and have proven efficiency in different chemical optimization contexts. Prominent examples cover **cluster structure optimization (CSO)** for atomic as well as molecular clusters and general parameter optimizations. These are, as encountered in this Thesis, both representatives of (hard) continuous optimization problems. Further, some discrete molecular design was also already addressed (*cf.* Section 3.1). Backed up with these insights and a general optimization framework for global optimization in chemistry that we have developed, namely OGOLEM, the goal was to transfer this expertise into the new regime of catalysis design.

A next pivotal step for this optimization generally is the shrinkage of the search space. Even though the aforementioned **EA** can lead to very efficient short-cuts in the property search space, prosperous approaches usually introduce either meaningful pre-generated libraries of molecules and/or further abstraction layers of the models used. Hence, this Thesis introduced the general concept of **globally optimal catalysts (GOCATs)** as such a model. These **GOCATs** consist of general and foremost abstract *interaction groups* placed nearby a chemical reaction, usually by surrounding it, for maximal reaction rate increase. This property, however, has to be defined in a thermodynamical and kinetical sense such that different objective functions were invented to best represent catalytic effects of this

kind. In the end, this presents a proper manipulation of the *effective potential energy surface (PES)* of the chemical systems. This is meant as a maximally reductionistic approach that allows to avoid tackling all the very subtle intricacies of real-life concrete catalysts at first and allows to focus on the underlying chemical interactions instead. The evident follow-up problem then arises to translate these *GOCATs* back to real functioning molecules. This second step was not addressed in this Thesis. As shown in many applications in this Thesis, however, a great portion of the understanding and predictions of such catalysis has already been possible without a translation.

As a meaningful starting point in this endeavor, *electrostatic GOCATs* that consist of a number of partial point charges were introduced. They are optimized around a reaction that is represented by many so-called frames. Naturally, these frames included the *reactant (R)*, the rate-determining first-order saddle-point, *i.e.*, the *transition state (TS)*, the *product (P)* as well as enough additional frames to create a quasi-continuous view on the *minimum energy path (MEP)* of the reaction. Mostly, all of these frames build up a superposed common surface for the partial charges such that the latter abstractly impose non-bonding interactions that are represented by Coulomb interactions around the reaction. The catalytic effect encoded in the objective function was based on the electronic energy barrier between *TS* and *R*, gradients at the stationary frames and more simplification ingredients to reach a balanced description. In applying such global optimization techniques, this really is one of the minimum, *i.e.* most simple, abstract models that should be used.

For instance, when just one *TS* frame and no other frames were the subject of pure energy-based minimization by *electrostatic GOCATs*, no optimizations would be needed in certain circumstances at all, but just chemical intuition and a pen and paper might be sufficient. This would, however, just mirror the molecular electrostatic potential of the *TS* structure and would lead to all types of artifacts for catalysis. For example, the *R* structure could also be stabilized by the exact same amount or maybe also completely destabilized. Further, maybe the *TS* could be no saddle point on the *PES*, *i.e.*, no *TS* anymore, if no gradient information were also included. As a consequence, a more balanced description that takes into account more objective function ingredients than just the energy and more frames than just the *TS* is essential. By generalizing this note, the problem of overfitting must be handled. In this context, overfitting generally means that already chemically less meaningful paths would be noticed afterwards if it were looked at more pieces or other types of information of the reaction than the utilized information during the optimization. Put differently, a seemingly well-performing solution that is estimated in this way by the objective function at first glance is not well-performing at second glance. Thus, the general findings with respect to these proper so-called fitness functions, *i.e.*, mostly the objective functions in the language of *EA*, were one central method-development result of this Thesis. This was discussed in Chapter 3.

Concluding, these *electrostatic GOCATs* tackle pure *electric field (EF)*-based catalysis for reactions. Electrostatics *can* be a great part of catalysis and, according to some authors, is considered to be the most important part. In fact, many recent studies both computationally as well as experimentally have focused also on pure *EF*-based catalysis (*cf.* the depictions

in Sections 6.3.3.1 and 7.1). Therefore, in this context no translations to real molecules from the abstract **GOCATs** are needed in order to already allow a useful inspection of the electrostatic effects. Due to the *global* optimization of **GOCATs**, this approach allows to get a glimpse of the maximal possibilities of **EF** catalysis. Since non-bonding interactions essentially are all based only on Coulomb terms, with enough of these Coulomb centers and a complete polarization, it should be possible to create or reach a maximal, potential catalytic effect, in this classical picture. Assuming this to be true, any other (non-**EF**) catalyst (for non-bonding interactions) would probably lead to less pronounced catalytic effects. Hence, with the electrostatic **GOCATs** one can gain an impression of what is maximally possible as an upper bound in any case. Even more so, by defining the partial point charges, a confidence in the global convergence of the optimization can be gained, which would not be possible if, *e.g.*, concrete molecular fragments were used. In other words, the search space has been shrunk, and the optimization problem has been simplified. Finally, having found the best **GOCATs** for a reaction, these can always serve as reference points for everything that could follow next in the catalysis design framework. This is true not only for all types of extensions to other interaction groups for **GOCAT** models but also when real molecular fragments are optimized. The globally optimal electrostatic field can even be a reference or immediate target for future translations to real molecular systems.

8.2.2 Summary of Observations

A big part of this Thesis dealt with method development of the **OGOLEM** suite. These extensions and improvements covered three different topics: parameter, cluster and the **GOCAT** optimization.

Parameter optimization: First, scaling improvements for the global parameter optimization for the reactive force field, **REAXFF**, were discussed. This was a side-project before tackling the **GOCAT** design task, and this partially goes back to the Master's Thesis of the current author. Linear strong scaling is an intrinsic feature of the **OGOLEM** suite in all other types of optimization problems because of our generation-free **EA** algorithm. Hence, this flaw with respect to the scaling for **REAXFF** optimization was readdressed and solved in this Thesis. This was possible with appropriate software designs on the **OGOLEM** side as well as low-end memory handling for the **REAXFF** backend. This side-project and the implementations were discussed in Chapter 4 and in particular in Section 4.3.

Cluster optimization: Second, in the regime of **CSO**, different so-called niching techniques were developed. In a more abstract sense, niching is the introduction of an order parameter into the **EA**-based search to conserve some intrinsic diversity with respect to the encoded chemical structures in the current **EA** pool of solutions. Without such a parameter, often final solutions prematurely converge, meaning that the global optimum in the search space has not yet been found. Since such a pool without niching could evolve to have no intrinsic diversity, no efficient progression towards the global optimum would be possible anymore. For this purpose, different **machine learning (ML)** descriptors of chemical matter were

visited. In this regard, **Lennard-Jones (LJ)** clusters were optimized that constitute a type of benchmark problem for hard **CSO**. Specific sizes of these clusters are especially cumbersome for the optimization because of very deceptive high-dimensional configurational energy surfaces such that the niching is pivotal for finding the (already known) global optima of these cluster sizes. In the end, both problem-specific descriptors based on sub-motifs in the atomic clusters and more global problem-independent descriptors were utilized. A **coulomb matrix (CM)** descriptor that belongs to the latter class showed to be of generally good performance for this niching as did also the problem-specific descriptor. One important finding was that such descriptors often do not have to be fine-tuned for the problem at hand. Rather the mere presence and only a rough meta-parameter benchmarking of such order parameters are sufficient and yet crucial to find the global optimum.

Since **CM** is an abstract descriptor, this can equally well be used for the **GOCAT** design. Thus, this project was also a touchstone for these specific niching descriptors as well as for the general niching algorithm implementation in **OGOLEM**. This was discussed in Chapter 5.

GOCAT optimization for the Menshutkin reaction: Third, the **GOCAT** theme was addressed for the optimal catalysis design. The basis of this task was developed from scratch and needed very many extensions to the **OGOLEM** suite. These included all types of different operators such as mutation, crossover and niching operators as well as underlying (local) optimization algorithms, reaction path optimizations, communication frameworks with other programs, data representations within **OGOLEM** and many more. Exactly these details make up the biggest part of this Thesis and are discussed in both Chapter 2 and Chapter 3.

Two applications of this **GOCAT** framework have made it into this Thesis at the end: (1) The electrostatic **GOCAT** optimization around a Menshutkin reaction and (2) an electrostatic **GOCAT** optimization around a **Diels–Alder (DA)** reaction.

(1) In Chapter 6, the S_N2 Menshutkin reaction as a proof-of-concept study was selected because catalytic effects follow a clear trend known *a priori*. The simple reasoning here is that this reaction starts with neutral reactants, goes over an already well polarized **TS** and finally reaches a contact ion pair when the products are formed. Any impact an electrostatic surrounding can have on the **R** structure will likely be enlarged for the **TS** because of the higher molecular dipole moment and the stronger polarization possibilities here, which results automatically in a higher dipole coupling between the **EF** and these molecules. Therefore, usual polar solvents do also stabilize the **P** structures, the **TS** structure and—to a smaller extend—also the **R** structures. As a result, the barrier is decreased and the **TS** is shifted on the **PES** more to the **R** side. These were the anticipations from the outset, and the optimized **GOCATs** were able to evolve this trend as a catalytic effect. This was also compared to usual implicit solvent models to contrast this fully global **GOCAT** model with pure polar embeddings that show such a catalytic effect in this case “by accident”.

The Menshutkin reaction was fully tackled in the *static* or *vertical* **GOCAT** mode. In this case, a pre-optimized gas phase reaction path for the reaction is to be stabilized by the **GOCAT**. No shifts of any structures on the **PES**, and consequently no reaction mechanism changes are allowed. Notwithstanding, clear electrostatic scalar potentials by the **GOCATs**

that optimize this reaction could be found. This study was a genuine benchmark of the whole **GOCAT** theme reaching from simple models starting with just *one* partial charge to more elaborated ones. With a certain size and flexibility of the **GOCAT**, which is encoded by the model restraints such as the inter-charge distance allowed, the bounds on the charge values and the reaction pocket surface, a *convergence* of the final maximal catalytic effect was also observed. This means that, with just enough flexibility of the model, the maximal electrostatic catalytic effect can be worked out already. An even more flexible **GOCAT** model would *not* improve the catalytic effect any further, but it would solely complicate the **GOCAT** solutions. Naturally, this **GOCAT** model convergence is dependent on the studied reaction. However, this illustrates an upper bound that could be found for the electrostatic catalytic effect for this case.

This points to a drawback of the framework that was often encountered. As a metaheuristic and non-deterministic global optimization was carried out, there are two fundamental reasons for noise effects in the final solutions space. First, it is never guaranteed by the metaheuristic approach that the global optimum really has been found after an optimization run. Yet, as this **GOCAT** design problem really has a huge and very intricate search space, there is no alternative to these metaheuristic approaches. Thus, this noise effect must be accepted. Second, due to the superposition of the Coulomb scalar potential anywhere around the reacting frames as well as shielding effects (due to both positive and negative charge values), there are many **GOCATs** that show an evidently different Cartesian shape and yet also generate a rather similar **electrostatic potential (ESP)**, at least at the important atom regions of the inner frames. This can be seen as growing linear dependency between the separate charges. As a result, there are many different solutions in the search space with almost the same fitness. The very best candidate solution for the global optimum, though, was a strikingly symmetric **GOCAT** without inducing any symmetry during the search. In combination of the two noise effects, this led to the necessity of using unsupervised **ML** techniques to compress the solution space into meaningful chemically discriminant representations as well as to utilize inferential statistical techniques to uncover the catalytic trends.

This same Chapter 6 about the Menshutkin reaction was supplemented with further Complementary studies. First in Section 6.3.1, an algorithmic benchmark was carried out for some of the implemented new **EA** operators. Here, the general finding was that a specific **ESP**-aware operator that introduced big steps through the search space, but reached a similar **ESP** afterwards, the so-called canada operator, showed the most effective **EA** moves. For specific optimization problems at hand, such as this concurrent **CSO** and parameter optimization problem, *i.e.* the **GOCAT** design, problem-specific operators can naturally outperform general black-box operators. Therefore, also other problem-specific operators were implemented. Finally, using a roughly benchmarked mixture of these problem-specific ones with some fraction of more problem-independent ones, has again shown to be the most robust operator setting.

Moreover, the translation between the levels of theory was discussed in Section 6.3.2. Because of the global optimization of the **GOCATs** which needs very many energy/gradient

calculations, a compromise with regard to the level of theory is often needed. To show a general qualitative translation between the used semi-empirical level of theory and a **density functional theory (DFT)** level of theory, different translation protocols were investigated. These ranged from mappings onto the new surface to local as well as global **ESP** optimizations. By changing the level of theory, the exposed surface of the reaction can also change due to the slightly different configurations of the structures. All translation protocols showed already catalytic **GOCATs** translated from a **PM7** semi-empirical treatment to a **PBE0 DFT** level without any further catalytic optimization for the Menshutkin reaction. In fact, the simpler methods without the local or even global **ESP** optimization performed better in this regard. This can probably be explained by the overfitting that is encountered yet another time when the **ESP** is globally optimized during the translation protocol, because of the increasing linear dependency of the charge centers. Here, more grid points defining the **ESP** in the reaction pocket should be used and perhaps some additional simplification ingredients to compress the solution during the translation.

Additionally, the mentioned very clear-cut trend observed in the Menshutkin reaction was re-addressed by simple linear regression and correlation analysis in Section 6.3.3. Due to the simplicity of this trend, it served as a reference point for the **DA** reaction. Furthermore, recent electrostatic catalysis research in other computational and experimental studies were set into perspective. The simple trend was to increase the **EF** of the **GOCAT** along the molecular dipole moment of the Menshutkin reaction without any further complexities of local dipole contributions or polarization effects that could otherwise change the dipole direction or lead to the necessity to include even higher moments in such interpretations.

GOCAT optimization for the Diels–Alder reaction: (2) In Chapter 7, the next subject of discussion was the **DA** reaction. This investigation used the improved *non-vertical* or *adaptive* fitness function for **GOCAT** design. A reference reaction path is still needed here, and this was again taken from the gas phase path, but in this case a complete **MEP** relaxation of all frames within the **GOCAT** is allowed. This enables bigger impacts of **GOCATs** on the **PES**, from small non-vertical energy shifts to complete mechanistic changes. The study was oriented along general computational and specific experimental **scanning tunneling microscope (STM)** results from the literature for the tackled **DA** reaction. These results have already proven that **EFs** can also catalyze non-polar non-redox reactions. With a stronger **EF**, generally, the polarization of the molecules increase such that even ionic states can be stabilized to form the most stable structure on the (adiabatic) **PES**. This trend emerged and was followed during the optimization by the unbiased electrostatic **GOCAT** approach.

Because of the two superposed noise effects mentioned above, intensive regression analyses and **ML** methods were carried out once more. Additionally, a completely uniform **EF**, which was always used in the literature studies, constituted a reference system for the **GOCAT**. In general, the trend of an increase of the **EF** along the so-called “reaction axis” (which is the dominating axis of electron density changes due to bond breaking or making) for the **DA** reaction induced the catalytic effect. With an increase of this field along the two C–C-bonds that are built in the concerted **DA** reaction, the catalytic effect increased.

Only after allowing the adaptive MEP relaxations, this catalytic effect increased so much that almost no barrier at all remained in the end. By this, the overall mechanism changed from the concerted symmetric DA reaction, over a concerted asymmetric reaction (with two different C–C-distances) to a two-step mechanism. In the latter case the polarization increased to such an extent that a zwitterionic intermediate is stable and found automatically during the GOCAT design.

These GOCATs were compared to other static GOCATs and the uniform EF models and showed again some increased complexities such as highly non-uniform field components by the GOCAT and also contributions in other directions than along the one reaction axis. This is to be expected since full polarization and a bunch of partial charges create all the sources and sinks for the ESP and the shielding effects of the Coulomb potential at certain Cartesian places.

Finally, it must be stressed that this solution complexity is two-fold: First, it does, indeed, complicate the simple interpretation of the catalytic effects. Second, there *are* many solutions in the search space that do not follow such simple trends, but still show a huge catalytic effect. By the ML techniques, *i.e.*, the insights provided by clustered representatives within the solution space and by the non-linear dimensionality reduction techniques, further GOCATs generating an unexpected ESP could readily be found and characterized. In order to be able to fully understand these candidate solutions, more research as well as comparisons with higher-level calculations are needed to exclude systematic errors of the semi-empirical treatment. However, even the clear-cut Menshutkin reaction on DFT level already showed such a solution variability in both the Cartesian *and* the ESP space. Thus, for a possible translation to real molecules or to understand certain electrostatic catalysis effects, this plethora of solutions could be an advantage when new, rather unexpected catalytic ESP domains can be utilized.

To conclude, the current author joins a quotation of SHAIK *et al.*:^[459]

[...] oriented external electric fields (OEEFs) [...] catalyse and control a variety of non-redox reactions and impart selectivity at will. An OEEF along the direction of electron reorganization [...] will catalyse nonpolar reactions by orders of magnitude, control regioselectivity and induce spin-state selectivity.

With the additional selected examples of the Master's Thesis of BEHRENS in Section 8.1, such controls “at will” by using a proper fitness function and the efficient global optimization always turned out to be successful in the end.

8.3 Prospects

In the following, some possible directions of further improvements and applications of the framework will be addressed. In fact, these thoughts are not just hypothetical since the GOCAT framework is already developed and expanded further by the successor of this topic, BEHRENS,^[489] at the very time of writing this Thesis.

First, future research with regard to the electrostatic **GOCAT** model, as used in this Thesis, could be carried out along the following lines:

- Because of the mentioned “noise” effects, many **GOCATs** seemingly represented just a minor version of better **GOCATs** within the same clusters. In any case, there was a high variability of different Cartesian and **ESP** domains in the solutions. These rather different solutions should be investigated further in order to understand non-intuitive electrostatic effects. To reproduce what is anticipated *a priori* is satisfactory with regard to proof-of-concept designs. When more notable applications are tackled, though, a prediction mode that also includes other subtle (not expected) effects would be interesting. In the perspective of SHAIK *et al.*,^[268,459,465] electrostatic catalysis with electric fields could “at will” also induce all types of chemical selectivities, even enantioselectivity. In the cited literature, they already mentioned “superposed” fields that could be applied along two directions, but that are still uniform **EFs**. With the **GOCAT** at hand, this “will” could simply be encoded into the fitness function and optimized right away.
- In the adaptive version of the **GOCAT**, mathematical graph-theory as a common descriptor of chemical molecules (since this simply mirrors Lewis’ chemical formulae, at least for common organic chemistry) should be implemented in order to have a control over this evolving **MEP** information during the global optimization. Different reaction paths could then be discriminated by using some niching technique, and, furthermore, different paths could be enforced as well. Taking an example from this Thesis, this could lead to an optimization of *endo* vs. *exo* **DA** reactions in *one* **EA** pool to directly induce the *kinetic* control of the diastereomeric outcome. Moreover, the **nudged elastic band (NEB)** improvements discussed in Section 2.5.1.4 could then be considered to increase the robustness of the algorithm even further.
- Additionally, simplification protocols should be included for the electrostatic **GOCATs**. With respect to Occam’s razor as heuristic for the catalysis design one could argue that if two **GOCAT** models are inducing a catalytic effect equally well, but one is apparently simpler than the other, the simpler one should be favored. On the one hand, either just the model restrictions could be tightened, such as the restriction to very small charge values, a smaller number of charges and overall less flexibility of the embedding due to higher inter-charge distances, or, on the other hand, penalties as regularization terms on such values could be included into the fitness function directly. For instance, if two charges with very different values (including the signs) are placed in the immediate neighborhood and/or a “cloud” of charges with high variance at specific sites, this regularization could penalize such needlessly complex candidate solutions. Here, different charges rather shield each other with only small effects on the resulting **ESP** such that these more complex **GOCATs** could be exterminated during the **EA** in this way. A controllable trade-off could be reached between maximally flexible but hard to interpret charge embeddings on the one hand and more smoothly varying smaller effects that can be interpreted (or translated)

more easily on the other hand. Note that such simpler solutions *are* already part of the search space used in this Thesis. However, these are not specifically favored by additional objective function ingredients. In other words, no fundamental new program features but mainly adaptations of the objective function were needed for charge complexity reduction.

- With respect to the goal of electrostatic catalysis *per se*, electrostatic **GOCATs** with and without theozymes for such systems as shown for Kemp eliminase in Section 8.1 could be further investigated. The authors of Refs. [123, 126] concluded that the electrostatics of the outer scaffold should be incorporated already during first stages of the *ne novo* design. With such an electrostatic **GOCAT**, the maximally possible improvements of such a full enzyme around the theozyme can be assessed. Furthermore, the **ESP** could as well be used as a guidance for these *de novo* design steps. This was already a point in the outlook of BEHRENS' Master's Thesis.

Next, general and fundamental **GOCAT** model approximations were already discussed in the publication in Section 6.2 on p. 142 (*cf.* Table 2 and the main text there). Indeed, the first three entries of the Table were already tackled and solved in this Thesis. They were the items “fixed/preoptimized reaction path”, “single step mechanism” and “given/fixed mechanism”, while these entries are partially overlapping. Some more future improvements and remarks for these three and the other entries of the Table can be outlined as follows:

- Having established a fully electrostatic reference with the **GOCATs** of this Thesis, the next step could be the development of small-molecule fragments as a discrete library of chemical compounds. When, *e.g.*, different amino acids (in also different protonation and deprotonation states) as well as some solvent molecules are provided, a *globally* optimal theozyme as **GOCAT** could be reached. With the incorporation of more frames than just the **TS** frame, also trends in stabilization *vs.* destabilization, themes of so-called “near attack-conformers” and maybe other discussed effects could be investigated for enzyme catalysis.^[70] In fact, the incorporation of small-molecule fragments *is* the next step taken by BEHRENS, which might be more important than other abstract model interactions such as **vdW** groups, H-bond centers, force centers (*cf.* Chapter 1).
- This fragment list could naturally also include typical ligands for transition-metal catalysis. With a tiny change of the objective function that is usually used (*i.e.*, with the change of some weights), this could—to the knowledge of the current author—be identical to the **GdMC** approach by WEYMUTH and REIHER,^[7,134,135] however, a global optimization would be used here. Arguably, by a clear shell-wise construction with more important inner and less important outer ligand shells, this chemical priority directly mirrors the greedy approach taken in the literature. Thus, a full-blown global optimization could lead to a (too) high computational burden in practice for this specific application.

Another question is whether *one* static surrounding of ligands for such a metal center suffices or whether an extension to *frame-wise* changing **GOCATs** is already

needed (see a point below). Without this extension, such **GOCATs** might influence first and foremost only one frame, the **TS**. In order to handle multiple frames and their relational energetic information, *i.e.*, a full **MEP**, ligand structures/positions, which belong to the **GOCAT**, should vary between the frames if this is crucial for the problem at hand.

- For transition-metal systems with clearly set ligand positions around the metal center, *discrete* spaces could be implemented. Around metal centers, the coordination configurations are often known, meaning that at least some priorities of which discrete sites are coordinated and which ligand angles (coordinating atoms) are involved can be specified. With discrete spaces, ligand positions could thus be better represented. Hence, incorporation of such information in a *relational* sense could shrink the search space and make the candidate solutions more chemically feasible. Reaching back to the discrete molecular design of azobenzenes in **OGOLEM**^[177] (this was briefly discussed in Chapter 3), similar features could be deployed that are already available in **OGOLEM** but that would need to be generalized.
- In the past, we have already optimized reactions within globally optimal solvation shells.^[400] Back then, pure energy minimizations were carried out for each frame of a reaction to generate the best (lowest energy) **MEP** possible in combined explicit and implicit water. However, each frame was optimized *independently* from the others. This could be pictured as a type of adiabatic separation between the reaction frames (**MEP**) and the surrounding water clusters, where the latter clusters are able to “infinitely” quickly adapt to each (separate) reaction frame. As a result, each separate neighbored pair of water clusters around the frames did not show a clear Cartesian similarity. Thus, this led to *one* extreme or corner case of the description of the solvent influence onto the reaction.

By contrast, the *other* extreme of the description in terms of “infinitely” fast reaction frames (*e.g.*, simply the same frames as before or relaxed ones) and essentially frozen (static) surroundings could be reached by the **GOCAT** model. At the moment and without further extensions yet, still *one* **GOCAT** would surround all the frames at once. Thus, as the other corner case of the description, this would lead to a complementary picture. When some small-molecule fragments are included into the **GOCAT**, such solvation shells can be readdressed from this perspective.

By defining a proper objective function, different effects such as a maximal **TS** stabilization and/or **R** destabilization could be investigated. When these results were compared to the complementary frame-wise changing water clusters of Ref. [400], catalytic and/or energetic effects could be analyzed. Different settings could be used: If a focus were placed on the **TS** structure, this would lead to an “anticipation” of the **TS** in a preorganized shell. In contrast, in equilibrated solvents, the stable (minima) structures of the solute determine the alignment of the outer solvent molecules. Hence, with focus on the **R** structure, a non-preorganized shell as a solution cluster that orients along the non-rare events could be optimized as another limiting case. All

these cases could be compared to the upper bound of maximal electrostatic catalysis by the partial charge **GOCATs**.

- One of the general further model restrictions so far, mentioned above multiple times already, is the *one* surrounding **GOCAT** for all frames at once. This was one important point in the Table of model restrictions discussed in the publication on p. 142. Instead of just *one* **GOCAT** for all frames combined, a frame-wise changing **GOCAT** could be implemented. This should be accompanied by an abstract distance criterion between these outer frames of the **GOCAT** to induce an upper bound of the maximum allowed change from frame to frame in order not to create sudden erratic “jumps” of **GOCAT** frames.

Assuming, for the sake of argument, that even if the global optimum with respect to the energy were found for each separate frame and this optimum were a discriminated one without any other equally-fit competitors, one could argue whether such surrounding **GOCATs** would already smoothly follow the reacting frames without any further “kinks”, at least in the limit of infinitely many frames (or a continuous **MEP**). In practice, all types of problems can occur such as a differing convergence behavior between the **GOCATs** at frame i and $i + 1$ when maybe the first **GOCAT** happens to stabilize a different mechanism or electronic state compared with the neighbor **GOCAT**,³ for instance. (This is similar to, *e.g.*, performing relaxed scans along specific coordinates in usual local energy minimizations of molecules, where the other maybe very loose coordinates might converge structurally differently between the separate restrained/constrained coordinates leading to “kinks” on the scanned **PES**.)

In any case, the global optimum is never guaranteed by the metaheuristic optimization, and additional distance measures enforcing a smooth chain of **GOCAT** frames should therefore be incorporated from the outset. By allowing **GOCATs** to change from frame to frame of the inner reaction, the possible catalytic effect will increase as now *no* compromise between different structures of the **PES** is imposed anymore. A correlated outer **GOCAT** “motion” with regard to the inner reaction would then be describable.

- After implementing the *frame-wise* **GOCATs** as well as the *discrete* spatial positioning of the **GOCAT** centers, transition-metal complexes (as well as theozymes) with the full-blown objective functions could be dealt with. This could include barriers of several steps in the cycle and could even relate multiple intermediates and **TS** structures differently. When manifold parallel and consecutive paths were attainable, it would be interesting to control such relational energetic effects with different **GOCATs**. Probably, this is one of the more distant ultimate improvements.

However, two further drawbacks must be noticed here. First, when a tight pocket

³ At the moment, always one electronic state, *i.e.*, one fixed electron number and spin state, is enforced during the whole reaction. But with the explorative nature of such global optimizations, it would also be imaginable that other states that are excluded by this manual selection beforehand might become the stable ones and that these are also open for optimization.

created by a common exposed surface of the reaction frames is used for the **GOCAT** positioning, all these evolved **GOCATs** and the reaction are held in a tightest associated state. In the bio-catalysis regime, one would speak of the enzyme–substrate complex. Hence, complete catalytic cycles of binding to the **GOCAT** and dissociating from the **GOCAT** afterwards can not be modeled in this way, at least if one presumes something as this closed surface. If more expanded surfaces for the **GOCAT** were used, such as the spherical ones with a lot of internal space, the internal reaction (including multiple different steps) could be described, maybe even with frame-wise changing **GOCATs** at the end. Another interesting question in connection with the last point would be whether a frame-wise changing **GOCAT** could be *evolved* to develop something as a “pocket” that is open or can be opened during the reaction to bind and release the substrate(s) and product(s), respectively, after additional elaborated changes to the fitness function. Second, even if the adaptive relaxed **MEP** version is used, some starting path must be chosen, which leads to the next item of this list.

- Another possible bias from the user could be the reference path the **GOCAT** optimization needs as input. These starting paths were mostly taken from the gas phase, but also other arbitrary ones are possible such that an unstable path can be stabilized (*cf.* the **COSMO** case of Section 6.2). For simple reactions, chemical intuition and manual **MEP** optimizations might suffice. Selecting a proper reference path for reactions (networks) that can be *very* complicated^[315,490] is then one further challenge for the computational chemist. MARTÍNEZ recently named this one of the “unknown unknowns” for reactions^[491] and referred to the nano-reactor^[233] as one possible exploration approach and to further automatic reaction network protocols.^[298,312–316,490,492,493]

The general problem is that in more complex reactions with multiple steps and involved species and maybe further needed model simplifications fundamental aspects could be overlooked without noticing. In the context of these intricate reaction networks, important other mechanistic steps, alternative paths and even other electronic states (spin) could escape the attention. Hence, such (semi-)automatic frameworks for studying the reaction(s) under consideration should be utilized. Different classes of such current approaches along these lines were recently reviewed in Ref. [494]. In the end, such methods should at least serve as orientation to decide on rate-limiting steps and which path(s) to start from. With such orientations, provision against overlooking important alternatives could be made. Put differently, opening “Pandora’s box” and allowing “every” combinatorial chemical possibility leads to the vast chemical space again that was obviated by the **GOCAT** model in the first place. As antidotes against dragging in the vast chemical space, one could limit it by suitable restrictions stemming from the *real* application (in the laboratory). Moreover, one could start again with a simpler setting with rigid restrictions on possible intermediates, **GOCAT** entities, on allowed variance of optimized paths from the starting paths and, in turn,

successively increase these limits until appropriate low-energy (catalyzed) reaction paths are found.

- Finally, the translation to real molecules could be tackled. One could sufficiently increase the library of possible small molecules and try to optimize these in a molecular assembly type of approach. Indeed, research along similar lines was already successful for other properties.^[495–498] Besides, such molecular assembly to reach pre-defined properties is already being developed in our work group, but it has not yet been published.^[499] When this project reaches a state in which it is able to assemble stable molecular systems or embeddings for certain pre-defined properties, these molecular assemblies could be oriented along the optimized (electrostatic) **GOCATs**.

Of course, partial charges in chemical systems are ill-posed, *i.e.*, simple monopole expansions of Coulomb centers representing the “fuzzy” electrons and static point-like nuclei cannot represent the (observable) electrostatic potential exactly, and, furthermore, common techniques for population analyses are not unique. Also in Ref. [7], different charge estimation schemes were utilized to translate abstract point charges on ligand positions rather “manually” back to atomistic ligands with similar charge properties. However, one could just ignore the exact Cartesian Coulomb charge centers in the electrostatic **GOCAT** (or rather use them as a first starting point only), and focus more on the actual **ESP** the charge centers create. The relevance of optimizing partial point charges of **GOCATs** in the first place is that more chemically “meaningful” scalar **ESP** values can be reached than by possibly tackling the optimization of an **ESP** vector without sources and sinks in absolute space. In other words, optimizing charge centers leads to a physically sound **ESP**, whereas optimizing an (even more abstract) **ESP** function—as **GOCAT**—directly would blow up the search space with (many) meaningless results. Indeed, often charges in electrostatic **GOCATs** tended to shield each other sometimes, and in other cases, they even created some evident quadrupole-like shape at specific sites as they, for instance, consisted of four charges with equal absolute value but inverse sign at opposite sites. This last point shall just emphasize the fact that partial point charges do not have to be mapped directly to partially screened real atoms at the *exact* same places of those partial charges before. Instead, one could first of all try to optimize/assembly real molecules around the reaction pocket (with temporarily absent reacting frames) in a type of least-squares optimization to accomplish the **ESP** at that surface points (*e.g.*, similar to the **vdW** surface pictures with mapped **ESP** values on top as shown in Figs. 7.4 and 7.11 on p. 187 and on p. 197, etc.). With the spectrum of *qualitatively* different **ESP** domains of **GOCATs** which lead so similar catalytic effects, one is not limited to only one **ESP** realization but could tackle multiple different ones and investigate which works best. As discussed in Chapter 7, having different possibilities for such a realization could be seen as a clear advantage then.

Furthermore, the abstract optimal **ESP** could serve as an orientation for the maximal catalytic effect, for the molecular assembly design and for further re-optimizations to reduce translation errors that will come to pass anyways. In later stages, a similar

objective function as used in this Thesis could serve for these reoptimizations of real concrete molecules to enhance the catalytic effect after the translation again.

Bibliography

- [1] M. Jansen, J. C. Schön, “DESIGN” IN CHEMICAL SYNTHESIS—AN ILLUSION?, *Angew. Chem. Int. Ed.* **2006**, *45*, 3406–3412 (cited on p. 1).
- [2] A. Aspuru-Guzik, R. Lindh, M. Reiher, THE MATTER SIMULATION (R)EVOLUTION, *ACS Cent. Sci.* **2018**, *4*, 144–152 (cited on p. 1).
- [3] P. A. M. Dirac, QUANTUM MECHANICS OF MANY-ELECTRON SYSTEMS, *Proc. R. Soc. London, Ser. A* **1929**, *123*, 714–733 (cited on p. 1).
- [4] C. Kuhn, D. N. Beratan, INVERSE STRATEGIES FOR MOLECULAR DESIGN, *J. Phys. Chem.* **1996**, *100*, 10595–10599 (cited on p. 1).
- [5] A. Jain, J. A. Bollinger, T. M. Truskett, INVERSE METHODS FOR MATERIAL DESIGN, *AIChE J.* **2014**, *60*, 2732–2740 (cited on pp. 1, 2).
- [6] O. A. von Lilienfeld, TOWARDS THE COMPUTATIONAL DESIGN OF COMPOUNDS FROM FIRST PRINCIPLES, in *Many-Electron Approaches in Physics, Chemistry and Mathematics: A Multidisciplinary View*, (Eds.: V. Bach, L. Delle Site), *Mathematical Physics Studies*, Springer International Publishing, Cham, **2014**, pp. 169–189 (cited on pp. 1, 2).
- [7] T. Weymuth, M. Reiher, INVERSE QUANTUM CHEMISTRY: CONCEPTS AND STRATEGIES FOR RATIONAL COMPOUND DESIGN, *Int. J. Quantum Chem.* **2014**, *114*, 823–837 (cited on pp. 1, 2, 8, 228, 232).
- [8] A. Zunger, INVERSE DESIGN IN SEARCH OF MATERIALS WITH TARGET FUNCTIONALITIES, *Nat. Rev. Chem.* **2018**, *2*, 0121 (cited on pp. 1, 2).
- [9] P. C. Sabatier, PAST AND FUTURE OF INVERSE PROBLEMS, *J. Math. Phys.* **2000**, *41*, 4082–4124 (cited on p. 1).
- [10] V. Ambarzumian, ÜBER EINE FRAGE DER EIGENWERTTHEORIE, *Z. Phys.* **1929**, *53*, 690–695 (cited on p. 2).
- [11] G. Borg, EINE UMKEHRUNG DER STURM-LIOUVILLESCHEN EIGENWERTAUFGABE: BESTIMMUNG DER DIFFERENTIALGLEICHUNG DURCH DIE EIGENWERTE, *Acta Math.* **1946**, *78*, 1 (cited on p. 2).
- [12] A. Donovan, H. Rabitz, EXPLORING THE HAMILTONIAN INVERSION LANDSCAPE, *Phys. Chem. Chem. Phys.* **2014**, *16*, 15615 (cited on p. 2).
- [13] K. W. Moore, A. Pechen, X.-J. Feng, J. Dominy, V. Beltrani, H. Rabitz, UNIVERSAL CHARACTERISTICS OF CHEMICAL SYNTHESIS AND PROPERTY OPTIMIZATION, *Chem. Sci.* **2011**, *2*, 417 (cited on p. 2).
- [14] K. W. Moore, A. Pechen, X.-J. Feng, J. Dominy, V. J. Beltrani, H. Rabitz, WHY IS CHEMICAL SYNTHESIS AND PROPERTY OPTIMIZATION EASIER THAN EXPECTED?, *Phys. Chem. Chem. Phys.* **2011**, *13*, 10048–10070 (cited on p. 2).
- [15] P. Kirkpatrick, C. Ellis, CHEMICAL SPACE, *Nature* **2004**, *432*, 823–823 (cited on p. 2).

- [16] A. M. Virshup, J. Contreras-Garcia, P. Wipf, W. Yang, D. N. Beratan, STOCHASTIC VOYAGES INTO UNCHARTED CHEMICAL SPACE PRODUCE A REPRESENTATIVE LIBRARY OF ALL POSSIBLE DRUG-LIKE COMPOUNDS, *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303 (cited on p. 2).
- [17] A. Mullard, THE DRUG-MAKER'S GUIDE TO THE GALAXY, *Nat. News* **2017**, *549*, 445 (cited on p. 2).
- [18] O. A. von Lilienfeld, FIRST PRINCIPLES VIEW ON CHEMICAL COMPOUND SPACE: GAINING RIGOROUS ATOMISTIC CONTROL OF MOLECULAR PROPERTIES, *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689 (cited on p. 2).
- [19] J.-L. Reymond, THE CHEMICAL SPACE PROJECT, *Acc. Chem. Res.* **2015**, *48*, 722–730 (cited on p. 2).
- [20] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, Alexandre Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, MACHINE LEARNING OF MOLECULAR ELECTRONIC PROPERTIES IN CHEMICAL COMPOUND SPACE, *New J. Phys.* **2013**, *15*, 095003 (cited on p. 2).
- [21] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, QUANTUM CHEMISTRY STRUCTURES AND PROPERTIES OF 134 KILO MOLECULES, *Sci. Data* **2014**, *1*, 140022 (cited on p. 2).
- [22] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, WHAT IS HIGH-THROUGHPUT VIRTUAL SCREENING? A PERSPECTIVE FROM ORGANIC MATERIALS DISCOVERY, *Annu. Rev. Mater. Res.* **2015**, *45*, 195–216 (cited on p. 2).
- [23] S. Hoelder, P. A. Clarke, P. Workman, DISCOVERY OF SMALL MOLECULE CANCER DRUGS: SUCCESSES, CHALLENGES AND OPPORTUNITIES, *Mol. Oncol.* **2012**, *6*, 155–176 (cited on p. 2).
- [24] B. Sanchez-Lengeling, A. Aspuru-Guzik, INVERSE MOLECULAR DESIGN USING MACHINE LEARNING: GENERATIVE MODELS FOR MATTER ENGINEERING, *Science* **2018**, *361*, 360–365 (cited on pp. 2, 4).
- [25] K. Sörensen, METAHEURISTICS—THE METAPHOR EXPOSED, *Intl. Trans. in Op. Res.* **2015**, *22*, 3–18 (cited on pp. 2, 24, 26).
- [26] K. Sörensen, M. Sevaux, F. Glover, A HISTORY OF METAHEURISTICS, **2017**, arXiv: 1704.00853 (cited on pp. 2, 24, 25).
- [27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, EQUATION OF STATE CALCULATIONS BY FAST COMPUTING MACHINES, *J. Chem. Phys.* **1953**, *21*, 1087–1092 (cited on pp. 2, 26).
- [28] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, OPTIMIZATION BY SIMULATED ANNEALING, *Science* **1983**, *220*, 671–680 (cited on pp. 2, 26).
- [29] T. Weise, GLOBAL OPTIMIZATION ALGORITHMS – THEORY AND APPLICATION, 3rd edition, www.it-weise.de, **2011** (cited on pp. 2, 13, 21–23, 25, 27–29, 70, 73, 74, 78, 85, 147, 149, 188).
- [30] D. E. Goldberg, GENETIC ALGORITHMS IN SEARCH, OPTIMIZATION, AND MACHINE LEARNING, 13th ed., Addison Wesley, Reading, Mass, **1989** (cited on pp. 2, 21, 28, 29).
- [31] D. E. Clark, D. R. Westhead, EVOLUTIONARY ALGORITHMS IN COMPUTER-AIDED MOLECULAR DESIGN, *J. Comput. Aided Mol. Des.* **1996**, *10*, 337–358 (cited on p. 2).
- [32] D. Xiao, I. Warnke, J. Bedford, V. S. Batista, CHAPTER 1: INVERSE MOLECULAR DESIGN FOR MATERIALS DISCOVERY, in *Chemical Modelling*, Vol. 10, The Royal Society of Chemistry, **2013**, pp. 1–31 (cited on p. 2).
- [33] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, A. Aspuru-Guzik, OPTIMIZING DISTRIBUTIONS OVER MOLECULAR SPACE. AN OBJECTIVE-REINFORCED GENERATIVE ADVERSARIAL NETWORK FOR INVERSE-DESIGN CHEMISTRY (ORGANIC), *chemrxiv:5309668* **2017** (cited on pp. 2, 4).

- [34] J. G. Freeze, H. R. Kelly, V. S. Batista, SEARCH FOR CATALYSTS BY INVERSE DESIGN: ARTIFICIAL INTELLIGENCE, MOUNTAIN CLIMBERS, AND ALCHEMISTS, *Chem. Rev.* **2019** (cited on pp. 2, 4).
- [35] M. Wang, X. Hu, D. N. Beratan, W. Yang, DESIGNING MOLECULES BY OPTIMIZING POTENTIALS, *J. Am. Chem. Soc.* **2006**, *128*, 3228–3232 (cited on p. 3).
- [36] O. A. von Lilienfeld, R. D. Lins, U. Rothlisberger, VARIATIONAL PARTICLE NUMBER APPROACH FOR RATIONAL COMPOUND DESIGN, *Phys. Rev. Lett.* **2005**, *95*, 153002 (cited on p. 3).
- [37] O. A. von Lilienfeld, M. E. Tuckerman, MOLECULAR GRAND-CANONICAL ENSEMBLE DENSITY FUNCTIONAL THEORY AND EXPLORATION OF CHEMICAL SPACE, *J. Chem. Phys.* **2006**, *125*, 154104 (cited on p. 3).
- [38] W. Yang, P. W. Ayers, Q. Wu, POTENTIAL FUNCTIONALS: DUAL TO DENSITY FUNCTIONALS AND SOLUTION TO THE v -REPRESENTABILITY PROBLEM, *Phys. Rev. Lett.* **2004**, *92*, 146404 (cited on p. 3).
- [39] S. Keinan, X. Hu, D. N. Beratan, W. Yang, DESIGNING MOLECULES WITH OPTIMAL PROPERTIES USING THE LINEAR COMBINATION OF ATOMIC POTENTIALS APPROACH IN AN AM1 SEMIEMPIRICAL FRAMEWORK, *J. Phys. Chem. A* **2007**, *111*, 176–181 (cited on p. 3).
- [40] D. Balamurugan, W. Yang, D. N. Beratan, EXPLORING CHEMICAL SPACE WITH DISCRETE, GRADIENT, AND HYBRID OPTIMIZATION METHODS, *J. Chem. Phys.* **2008**, *129*, 174105 (cited on p. 3).
- [41] X. Hu, D. N. Beratan, W. Yang, A GRADIENT-DIRECTED MONTE CARLO APPROACH TO MOLECULAR DESIGN, *J. Chem. Phys.* **2008**, *129*, 064102 (cited on p. 3).
- [42] S. Keinan, M. J. Therien, D. N. Beratan, W. Yang, MOLECULAR DESIGN OF PORPHYRIN-BASED NONLINEAR OPTICAL MATERIALS, *J. Phys. Chem. A* **2008**, *112*, 12203–12207 (cited on p. 3).
- [43] F. De Vleeschouwer, W. Yang, D. N. Beratan, P. Geerlings, F. De Proft, INVERSE DESIGN OF MOLECULES WITH OPTIMAL REACTIVITY PROPERTIES: ACIDITY OF 2-NAPHTHOL DERIVATIVES, *Phys. Chem. Chem. Phys.* **2012**, *14*, 16002 (cited on p. 3).
- [44] X. Hu, D. N. Beratan, W. Yang, A GRADIENT-DIRECTED MONTE CARLO METHOD FOR GLOBAL OPTIMIZATION IN A DISCRETE SPACE: APPLICATION TO PROTEIN SEQUENCE DESIGN AND FOLDING, *J. Chem. Phys.* **2009**, *131*, 154117 (cited on p. 3).
- [45] X. Hu, H. Hu, D. N. Beratan, W. Yang, A GRADIENT-DIRECTED MONTE CARLO APPROACH FOR PROTEIN DESIGN, *J. Comput. Chem.* **2010**, *31*, 2164–2168 (cited on p. 3).
- [46] D. Xiao, W. Yang, D. N. Beratan, INVERSE MOLECULAR DESIGN IN A TIGHT-BINDING FRAMEWORK, *J. Chem. Phys.* **2008**, *129*, 044106 (cited on p. 3).
- [47] D. Xiao, L. A. Martini, R. C. Snoeberger, R. H. Crabtree, V. S. Batista, INVERSE DESIGN AND SYNTHESIS OF ACAC-COUMARIN ANCHORS FOR ROBUST TiO₂ SENSITIZATION, *J. Am. Chem. Soc.* **2011**, *133*, 9014–9022 (cited on p. 3).
- [48] A. M. Chang, B. Rudshiteyn, I. Warnke, V. S. Batista, INVERSE DESIGN OF A CATALYST FOR AQUEOUS CO/CO₂ CONVERSION INFORMED BY THE Ni II-IMINOTHIOATE COMPLEX, *Inorg. Chem.* **2018** (cited on p. 3).
- [49] J. G. Kirkwood, STATISTICAL MECHANICS OF FLUID MIXTURES, *J. Chem. Phys.* **1935**, *3*, 300–313 (cited on p. 3).
- [50] O. A. von Lilienfeld, M. E. Tuckerman, ALCHEMICAL VARIATIONS OF INTERMOLECULAR ENERGIES ACCORDING TO MOLECULAR GRAND-CANONICAL ENSEMBLE DENSITY FUNCTIONAL THEORY, *J. Chem. Theory Comput.* **2007**, *3*, 1083–1090 (cited on p. 3).
- [51] V. Marcon, O. A. von Lilienfeld, D. Andrienko, TUNING ELECTRONIC EIGENVALUES OF BENZENE VIA DOPING, *J. Chem. Phys.* **2007**, *127*, 064305 (cited on p. 3).

- [52] O. A. von Lilienfeld, ACCURATE AB INITIO ENERGY GRADIENTS IN CHEMICAL COMPOUND SPACE, *J. Chem. Phys.* **2009**, *131*, 164102 (cited on p. 4).
- [53] D. Sheppard, G. Henkelman, O. A. von Lilienfeld, ALCHEMICAL DERIVATIVES OF REACTION ENERGETICS, *J. Chem. Phys.* **2010**, *133*, 084104 (cited on p. 4).
- [54] K. Saravanan, J. R. Kitchin, O. A. von Lilienfeld, J. A. Keith, ALCHEMICAL PREDICTIONS FOR COMPUTATIONAL CATALYSIS: POTENTIAL AND LIMITATIONS, *J. Phys. Chem. Lett.* **2017**, *8*, 5002–5007 (cited on p. 4).
- [55] B. Huang, O. A. von Lilienfeld, THE “DNA” OF CHEMISTRY: SCALABLE QUANTUM MACHINE LEARNING WITH “AMONS”, **2017**, arXiv: 1707.04146 (cited on p. 4).
- [56] R. Ramakrishnan, O. A. von Lilienfeld, MACHINE LEARNING, QUANTUM CHEMISTRY, AND CHEMICAL SPACE, in *Reviews in Computational Chemistry*, Vol. 30, (Eds.: A. L. Parrill, K. B. Lipkowitz), John Wiley & Sons, Ltd, **2017**, pp. 225–256 (cited on pp. 4, 62).
- [57] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakithodi, C. Kim, MACHINE LEARNING IN MATERIALS INFORMATICS: RECENT APPLICATIONS AND PROSPECTS, *Npj Comput. Mater.* **2017**, *3*, 54 (cited on pp. 4, 62).
- [58] B. Huang, N. O. Symonds, O. A. von Lilienfeld, QUANTUM MACHINE LEARNING IN CHEMISTRY AND MATERIALS, in *Handbook of Materials Modeling: Methods: Theory and Modeling*, (Eds.: W. Andreoni, S. Yip), Springer International Publishing, Cham, **2018**, pp. 1–27 (cited on pp. 4, 62).
- [59] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, AUTOMATIC CHEMICAL DESIGN USING A DATA-DRIVEN CONTINUOUS REPRESENTATION OF MOLECULES, *ACS Cent. Sci.* **2018**, *4*, 268–276 (cited on p. 4).
- [60] C. Poree, F. Schoenebeck, A HOLY GRAIL IN CHEMISTRY: COMPUTATIONAL CATALYST DESIGN: FEASIBLE OR FICTION?, *Acc. Chem. Res.* **2017**, *50*, 605–608 (cited on p. 4).
- [61] R. Wolfenden, M. J. Snider, THE DEPTH OF CHEMICAL TIME AND THE POWER OF ENZYMES AS CATALYSTS, *Acc. Chem. Res.* **2001**, *34*, 938–945 (cited on p. 4).
- [62] J. B. Beilen, Z. Li, ENZYME TECHNOLOGY: AN OVERVIEW, *Curr. Opin. Chem. Biol.* **2002**, *13*, 338–344 (cited on p. 4).
- [63] C. R. Catlow, M. Davidson, C. Hardacre, G. J. Hutchings, CATALYSIS MAKING THE WORLD A BETTER PLACE, *Phil. Trans. R. Soc. A* **2016**, *374*, 20150089 (cited on p. 4).
- [64] S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, M.-H. Baik, DESIGN AND OPTIMIZATION OF CATALYSTS BASED ON MECHANISTIC INSIGHTS DERIVED FROM QUANTUM CHEMICAL REACTION MODELING, *Chem. Rev.* **2019** (cited on p. 4).
- [65] S. Hammes-Schiffer, CATALYSTS BY DESIGN: THE POWER OF THEORY, *Acc. Chem. Res.* **2017**, *50*, 561–566 (cited on p. 4).
- [66] H. Eyring, THE ACTIVATED COMPLEX IN CHEMICAL REACTIONS, *J. Chem. Phys.* **1935**, *3*, 107–115 (cited on p. 5).
- [67] D. G. Truhlar, B. C. Garrett, S. J. Klippenstein, CURRENT STATUS OF TRANSITION-STATE THEORY, *J. Phys. Chem.* **1996**, *100*, 12771–12800 (cited on p. 5).
- [68] K. J. Laidler, CHEMICAL KINETICS, 3rd edition, Harper & Row, New York, **1987** (cited on p. 5).
- [69] J. Gao, S. Ma, D. T. Major, K. Nam, J. Pu, D. G. Truhlar, MECHANISMS AND FREE ENERGIES OF ENZYMATIC REACTIONS, *Chem. Rev.* **2006**, *106*, 3188–3209 (cited on pp. 5, 7, 178).

- [70] M. Garcia-Viloca, J. Gao, M. Karplus, D. G. Truhlar, HOW ENZYMES WORK: ANALYSIS BY MODERN RATE THEORY AND COMPUTER SIMULATIONS, *Science* **2004**, *303*, 186–195 (cited on pp. 5, 228).
- [71] J. K. Nørskov, T. Bligaard, J. Rossmeisl, C. H. Christensen, TOWARDS THE COMPUTATIONAL DESIGN OF SOLID CATALYSTS, *Nat. Chem.* **2009**, *1*, 37 (cited on p. 5).
- [72] C. H. Christensen, J. K. Nørskov, A MOLECULAR VIEW OF HETEROGENEOUS CATALYSIS, *J. Chem. Phys.* **2008**, *128*, 182503 (cited on p. 5).
- [73] L. Pauling, MOLECULAR ARCHITECTURE AND BIOLOGICAL REACTIONS, *Chem. Eng. News* **1946**, *24*, 1375–1377 (cited on p. 5).
- [74] L. Pauling, NATURE OF FORCES BETWEEN LARGE MOLECULES OF BIOLOGICAL INTEREST, *Nature* **1948**, *161*, 707–709 (cited on p. 5).
- [75] R. Wolfenden, TRANSITION STATE ANALOGUES FOR ENZYME CATALYSIS, *Nature* **1969**, *223*, 704–705 (cited on p. 5).
- [76] J. Kraut, HOW DO ENZYMES WORK?, *Science* **1988**, *242*, 533–540 (cited on p. 5).
- [77] W. A. Sokalski, THE PHYSICAL NATURE OF CATALYTIC ACTIVITY DUE TO THE MOLECULAR ENVIRONMENT IN TERMS OF INTERMOLECULAR INTERACTION THEORY: DERIVATION OF SIMPLIFIED MODELS, *J. Mol. Catalysis* **1985**, *30*, 395–410 (cited on p. 5).
- [78] W. A. Sokalski, NONEMPIRICAL MODELING OF THE STATIC AND DYNAMIC PROPERTIES OF THE OPTIMUM ENVIRONMENT FOR CHEMICAL REACTIONS, *J. Mol. Struct. THEOCHEM* **1986**, *138*, 77–87 (cited on p. 5).
- [79] W. A. Sokalski, THEORETICAL MODEL FOR EXPLORATION OF CATALYTIC ACTIVITY OF ENZYMES AND DESIGN OF NEW CATALYSTS: CO₂ HYDRATION REACTION, *Int. J. Quantum Chem.* **1981**, *20*, 231–240 (cited on pp. 5, 76).
- [80] B. Szeferczyk, A. J. Mulholland, K. E. Ranaghan, W. A. Sokalski, DIFFERENTIAL TRANSITION-STATE STABILIZATION IN ENZYME CATALYSIS: QUANTUM CHEMICAL ANALYSIS OF INTERACTIONS IN THE CHORISMATE MUTASE REACTION AND PREDICTION OF THE OPTIMAL CATALYTIC FIELD, *J. Am. Chem. Soc.* **2004**, *126*, 16148–16159 (cited on pp. 5, 6).
- [81] W. A. Sokalski, S. Roszak, K. Pecul, AN EFFICIENT PROCEDURE FOR DECOMPOSITION OF THE SCF INTERACTION ENERGY INTO COMPONENTS WITH REDUCED BASIS SET DEPENDENCE, *Chem. Phys. Lett.* **1988**, *153*, 153–159 (cited on p. 6).
- [82] P. Dziekonski, W. A. Sokalski, J. Leszczynski, PHYSICAL NATURE OF ENVIRONMENTAL EFFECTS ON INTERMOLECULAR PROTON TRANSFER IN (O₂NOH···NH₃)(H₂O)_N AND (CLH···NH₃)(H₂O)_N (*n*=1–3) COMPLEXES, *Chem. Phys.* **2001**, *272*, 37–45 (cited on p. 6).
- [83] W. A. Sokalski, R. W. Góra, W. Bartkowiak, P. Kobyliński, J. Sworakowski, A. Chyla, J. Leszczyński, NEW THEORETICAL INSIGHT INTO THE THERMAL CIS–TRANS ISOMERIZATION OF AZO COMPOUNDS: PROTONATION LOWERS THE ACTIVATION BARRIER, *J. Chem. Phys.* **2001**, *114*, 5504 (cited on p. 6).
- [84] P. Dziekoński, W. A. Sokalski, B. Szyja, J. Leszczynski, PHYSICAL NATURE OF CATALYTIC EFFECTS OF Si→Al SUBSTITUTIONS IN ZMS-5 ZEOLITE FOR PROPYLENE PROTONATION REACTION, *Chem. Phys. Lett.* **2002**, *364*, 133–138 (cited on p. 6).
- [85] P. Dziekonski, W. A. Sokalski, Y. Podolyan, J. Leszczynski, NONEMPIRICAL ANALYSIS OF THE CATALYTIC ACTIVITY OF THE MOLECULAR ENVIRONMENT – OPTIMAL STATIC AND DYNAMIC CATALYTIC FIELDS FOR DOUBLE PROTON TRANSFER IN FORMAMIDE–FORMAMIDINE COMPLEX, *Chem. Phys. Lett.* **2003**, *367*, 367–375 (cited on p. 6).

- [86] P. Kedzierski, P. Wielgus, A. Sikora, W. A. Sokalski, J. Leszczynski, VISUALIZATION OF THE DIFFERENTIAL TRANSITION STATE STABILIZATION WITHIN THE ACTIVE SITE ENVIRONMENT, *Int. J. Mol. Sci.* **2004**, *5*, 186–195 (cited on p. 6).
- [87] E. Dyguda-Kazimierowicz, W. Sokalski, J. Leszczyński, NON-EMPIRICAL STUDY OF THE PHOSPHORYLATION REACTION CATALYZED BY 4-METHYL-5- β -HYDROXYETHYLTHIAZOLE KINASE: RELEVANCE OF THE THEORY OF INTERMOLECULAR INTERACTIONS, *J. Mol. Model.* **2007**, *13*, 839–849 (cited on p. 6).
- [88] P. Szarek, E. Dyguda-Kazimierowicz, A. Tachibana, W. A. Sokalski, PHYSICAL NATURE OF INTERMOLECULAR INTERACTIONS WITHIN cAMP-DEPENDENT PROTEIN KINASE ACTIVE SITE: DIFFERENTIAL TRANSITION STATE STABILIZATION IN PHOSPHORYL TRANSFER REACTION, *J. Phys. Chem. B* **2008**, *112*, 11819–11826 (cited on p. 6).
- [89] E. I. Chudyk, E. Dyguda-Kazimierowicz, K. M. Langner, W. A. Sokalski, A. Lodola, M. Mor, J. Sirirak, A. J. Mulholland, NONEMPIRICAL ENERGETIC ANALYSIS OF REACTIVITY AND COVALENT INHIBITION OF FATTY ACID AMIDE HYDROLASE, *J. Phys. Chem. B* **2013**, *117*, 6656–6666 (cited on p. 6).
- [90] M. Chojnacka, M. Feliks, W. Beker, W. A. Sokalski, PREDICTING SUBSTITUENT EFFECTS ON ACTIVATION ENERGY CHANGES BY STATIC CATALYTIC FIELDS, *J. Mol. Model.* **2017**, *24*, 28 (cited on p. 6).
- [91] W. Beker, M. W. van der Kamp, A. J. Mulholland, W. A. Sokalski, RAPID ESTIMATION OF CATALYTIC EFFICIENCY BY CUMULATIVE ATOMIC MULTIPOLE MOMENTS: APPLICATION TO KETOSTEROID ISOMERASE MUTANTS, *J. Chem. Theory Comput.* **2017**, *13*, 945–955 (cited on p. 6).
- [92] C. K. Bagdassarian, V. L. Schramm, S. D. Schwartz, MOLECULAR ELECTROSTATIC POTENTIAL ANALYSIS FOR ENZYMATIC SUBSTRATES, COMPETITIVE INHIBITORS, AND TRANSITION-STATE INHIBITORS, *J. Am. Chem. Soc.* **1996**, *118*, 8825–8836 (cited on p. 6).
- [93] T. Gérczei, B. Asbóth, G. Náray-Szabó, CONSERVATIVE ELECTROSTATIC POTENTIAL PATTERNS AT ENZYME ACTIVE SITES: THE ANION–CATION–ANION TRIAD, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 310–315 (cited on p. 6).
- [94] E. Kangas, B. Tidor, ELECTROSTATIC COMPLEMENTARITY AT LIGAND BINDING SITES: APPLICATION TO CHORISMATE MUTASES, *J. Phys. Chem. B* **2001**, *105*, 880–888 (cited on p. 6).
- [95] M. Barbany, H. Gutiérrez-de-Terán, F. Sanz, J. Villà-Freixa, A. Warshel, ON THE GENERATION OF CATALYTIC ANTIBODIES BY TRANSITION STATE ANALOGUES, *ChemBioChem* **2003**, *4*, 277–285 (cited on p. 6).
- [96] A. Heine, E. A. J. T. Stura, Yli-Kauhaluoma, C. Gao, Q. Deng, B. R. Beno, K. N. Houk, K. D. Janda, I. A. Wilson, AN ANTIBODY EXO DIELS-ALDERASE INHIBITOR COMPLEX AT 1.95 ANGSTROM RESOLUTION, *Science* **1998**, *279*, 1934–1940 (cited on p. 6).
- [97] J. Na, K. Houk, PREDICTING ANTIBODY CATALYST SELECTIVITY FROM OPTIMUM BINDING OF CATALYTIC GROUPS TO A HAPTEN, *J. Am. Chem. Soc.* **1996**, *118*, 9204–9205 (cited on p. 6).
- [98] J. DeChancie, F. R. Clemente, A. J. Smith, H. Gunaydin, Y.-L. Zhao, X. Zhang, K. Houk, HOW SIMILAR ARE ENZYME ACTIVE SITE GEOMETRIES DERIVED FROM QUANTUM MECHANICAL THEOZYMES TO CRYSTAL STRUCTURES OF ENZYME-INHIBITOR COMPLEXES? IMPLICATIONS FOR ENZYME DESIGN, *Protein Sci.* **2007**, *16*, 1851–1866 (cited on p. 6).
- [99] D. J. Tantillo, C. Jiangang, K. N. Houk, THEOZYMES AND COMPUZYMES: THEORETICAL MODELS FOR BIOLOGICAL CATALYSIS, *Curr. Opin. Chem. Biol.* **1998**, *2*, 743–750 (cited on p. 6).
- [100] X. Zhang, J. DeChancie, H. Gunaydin, A. B. Chowdry, F. R. Clemente, Smith, T. M. Handel, K. N. Houk, QUANTUM MECHANICAL DESIGN OF ENZYME ACTIVE SITES, *J. Org. Chem.* **2008**, *73*, 889–899 (cited on p. 6).

- [101] B. I. Dahiyat, S. L. Mayo, DE NOVO PROTEIN DESIGN: FULLY AUTOMATED SEQUENCE SELECTION, *Science* **1997**, *278*, 82–87 (cited on p. 6).
- [102] B. I. Dahiyat, S. L. Mayo, PROTEIN DESIGN AUTOMATION, *Protein Sci.* **1996**, *5*, 895–903 (cited on p. 6).
- [103] A. Zanghellini, L. Jiang, A. M. Wollacott, G. Cheng, J. Meiler, E. A. Althoff, D. Röthlisberger, D. Baker, NEW ALGORITHMS AND AN IN SILICO BENCHMARK FOR COMPUTATIONAL ENZYME DESIGN, *Protein Sci.* **2006**, *15*, 2785–2794 (cited on p. 6).
- [104] G. Kiss, N. Çelebi-Ölçüm, R. Moretti, D. Baker, K. N. Houk, COMPUTATIONAL ENZYME DESIGN, *Angew. Chem. Int. Ed.* **2013**, *52*, 5700–5725 (cited on p. 7).
- [105] P.-S. Huang, S. E. Boyken, D. Baker, THE COMING OF AGE OF DE NOVO PROTEIN DESIGN, *Nature* **2016**, *537*, 320–327 (cited on p. 7).
- [106] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O’Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, THE ROSETTA ALL-ATOM ENERGY FUNCTION FOR MACROMOLECULAR MODELING AND DESIGN, *J. Chem. Theory Comput.* **2017**, *13*, 3031–3048 (cited on p. 7).
- [107] C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson, B. L. Stoddard, D. Baker, COMPUTATIONAL DESIGN OF LIGAND-BINDING PROTEINS WITH HIGH AFFINITY AND SELECTIVITY, *Nature* **2013**, *501*, 212 (cited on p. 7).
- [108] H. J. Wijma, D. B. Janssen, COMPUTATIONAL DESIGN GAINS MOMENTUM IN ENZYME CATALYSIS ENGINEERING, *FEBS J.* **2013**, *280*, 2948–2960 (cited on p. 7).
- [109] D. Hilvert, DESIGN OF PROTEIN CATALYSTS, *Annu. Rev. Biochem.* **2013**, *82*, 447–470 (cited on p. 7).
- [110] V. Nanda, R. L. Koder, DESIGNING ARTIFICIAL ENZYMES BY INTUITION AND COMPUTATION, *Nat. Chem.* **2010**, *2*, 15–24 (cited on p. 7).
- [111] D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, D. Baker, KEMP ELIMINATION CATALYSTS BY COMPUTATIONAL ENZYME DESIGN, *Nature* **2008**, *453*, 190 (cited on pp. 7, 214).
- [112] A. Warshel, ENERGETICS OF ENZYME CATALYSIS, *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 5250–5254 (cited on pp. 7, 172, 178).
- [113] A. Warshel, ELECTROSTATIC BASIS OF STRUCTURE-FUNCTION CORRELATION IN PROTEINS, *Acc. Chem. Res.* **1981**, *14*, 284–290 (cited on pp. 7, 178).
- [114] A. Warshel, ELECTROSTATIC ORIGIN OF THE CATALYTIC POWER OF ENZYMES AND THE ROLE OF PREORGANIZED ACTIVE SITES, *J. Biol. Chem.* **1998**, *273*, 27035–27038 (cited on pp. 7, 172, 178).
- [115] A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. Liu, M. H. M. Olsson, ELECTROSTATIC BASIS FOR ENZYME CATALYSIS, *Chem. Rev.* **2006**, *106*, 3210–3235 (cited on pp. 7, 172, 178).
- [116] S. C. L. Kamerlin, A. Warshel, AT THE DAWN OF THE 21ST CENTURY: IS DYNAMICS THE MISSING LINK FOR UNDERSTANDING ENZYME CATALYSIS?, *Proteins: Struct. Funct. Bioinf.* **2010**, *78*, 1339–1375 (cited on pp. 7, 172, 178).
- [117] A. Warshel, R. P. Bora, PERSPECTIVE: DEFINING AND QUANTIFYING THE ROLE OF DYNAMICS IN ENZYME CATALYSIS, *J. Chem. Phys.* **2016**, *144*, 180901 (cited on pp. 7, 172, 178).
- [118] A. Jiménez, P. Clapés, R. Crehuet, A DYNAMIC VIEW OF ENZYME CATALYSIS, *J. Mol. Model.* **2008**, *14*, 735–746 (cited on pp. 7, 178).

- [119] A. Christofferson, L. Zhao, Q. Pei, DYNAMIC SIMULATIONS AS A COMPLEMENT TO EXPERIMENTAL STUDIES OF ENZYME MECHANISMS, *Adv. Prot. Chem. Struct. Biol.* **2012**, 293–335 (cited on pp. 7, 178).
- [120] W. Lai, H. Chen, K.-B. Cho, S. Shaik, EXTERNAL ELECTRIC FIELD CAN CONTROL THE CATALYTIC CYCLE OF CYTOCHROME P450CAM: A QM/MM STUDY, *J. Phys. Chem. Lett.* **2010**, 1, 2082–2087 (cited on pp. 7, 178).
- [121] S. D. Fried, S. Bagchi, S. G. Boxer, EXTREME ELECTRIC FIELDS POWER CATALYSIS IN THE ACTIVE SITE OF KETOSTEROID ISOMERASE, *Science* **2014**, 346, 1510–1514 (cited on pp. 7, 171, 178, 179, 213, 214).
- [122] S. D. Fried, S. G. Boxer, ELECTRIC FIELDS AND ENZYME CATALYSIS, *Annu. Rev. Biochem.* **2017**, 86, 387–415 (cited on pp. 7, 171–173, 177–179, 183, 208, 213, 214).
- [123] A. Bhowmick, S. C. Sharma, T. Head-Gordon, THE IMPORTANCE OF THE SCAFFOLD FOR DE NOVO ENZYMES: A CASE STUDY WITH KEMP ELIMINASE, *J. Am. Chem. Soc.* **2017**, 139, 5793–5800 (cited on pp. 7, 171, 173, 178, 179, 183, 185, 214, 228).
- [124] V. V. Welborn, L. Ruiz Pestana, T. Head-Gordon, COMPUTATIONAL OPTIMIZATION OF ELECTRIC FIELDS FOR BETTER CATALYSIS DESIGN, *Nat. Catal.* **2018**, 1, 649–655 (cited on pp. 7, 171, 173, 178, 179, 183, 185).
- [125] V. V. Welborn, T. Head-Gordon, COMPUTATIONAL DESIGN OF SYNTHETIC ENZYMES, *Chem. Rev.* **2018** (cited on pp. 7, 178).
- [126] V. Vaissier, S. C. Sharma, K. Schaettle, T. Zhang, T. Head-Gordon, COMPUTATIONAL OPTIMIZATION OF ELECTRIC FIELDS FOR IMPROVING CATALYSIS OF A DESIGNED KEMP ELIMINASE, *ACS Catal.* **2018**, 8, 219–227 (cited on pp. 7, 171, 173, 178, 179, 183, 185, 214, 228).
- [127] A. Prah, E. Frančišković, J. Mavri, J. Stare, ELECTROSTATICS AS THE DRIVING FORCE BEHIND THE CATALYTIC FUNCTION OF THE MONOAMINE OXIDASE A ENZYME CONFIRMED BY QUANTUM COMPUTATIONS, *ACS Catal.* **2019**, 9, 1231–1240 (cited on pp. 7, 171, 173, 178, 179, 183, 185, 208).
- [128] V. V. Welborn, T. Head-Gordon, FLUCTUATIONS OF ELECTRIC FIELDS IN THE ACTIVE SITE OF THE ENZYME KETOSTEROID ISOMERASE, **2019**, arXiv: 1905.07521 (cited on pp. 7, 178).
- [129] P. Politzer, J. S. Murray, T. Clark, MATHEMATICAL MODELING AND PHYSICAL REALITY IN NONCOVALENT INTERACTIONS, *J. Mol. Model.* **2015**, 21 (cited on p. 7).
- [130] T. Clark, POLARIZATION, DONOR–ACCEPTOR INTERACTIONS, AND COVALENT CONTRIBUTIONS IN WEAK INTERACTIONS: A CLARIFICATION, *J. Mol. Model.* **2017**, 23, 297 (cited on p. 7).
- [131] T. Clark, HALOGEN BONDS AND σ -HOLES, *Faraday Discuss.* **2017**, 203, 9–27 (cited on p. 7).
- [132] T. Clark, A. Heßelmann, THE COULOMBIC σ -HOLE MODEL DESCRIBES BONDING IN $CX_3I \cdots Y^-$ COMPLEXES COMPLETELY, *Phys. Chem. Chem. Phys.* **2018** (cited on p. 7).
- [133] T. Clark, J. S. Murray, P. Politzer, A PERSPECTIVE ON QUANTUM MECHANICS AND CHEMICAL CONCEPTS IN DESCRIBING NONCOVALENT INTERACTIONS, *Phys. Chem. Chem. Phys.* **2018**, 20, 30076–30082 (cited on p. 7).
- [134] T. Weymuth, M. Reiher, GRADIENT-DRIVEN MOLECULE CONSTRUCTION: AN INVERSE APPROACH APPLIED TO THE DESIGN OF SMALL-MOLECULE FIXATING CATALYSTS, *Int. J. Quantum Chem.* **2014**, 114, 838–850 (cited on pp. 8, 228).
- [135] F. Krausbeck, J.-G. Sobez, M. Reiher, STABILIZATION OF ACTIVATED FRAGMENTS BY SHELL-WISE CONSTRUCTION OF AN EMBEDDING ENVIRONMENT, *J. Comput. Chem.* **2017**, 38, 1023–1038 (cited on pp. 8, 228).
- [136] Z. Zhou, X. Li, R. N. Zare, OPTIMIZING CHEMICAL REACTIONS WITH DEEP REINFORCEMENT LEARNING, *ACS Cent. Sci.* **2017**, 3, 1337–1344 (cited on p. 9).

- [137] S. Rangarajan, C. T. Maravelias, M. Mavrikakis, SEQUENTIAL-OPTIMIZATION-BASED FRAMEWORK FOR ROBUST MODELING AND DESIGN OF HETEROGENEOUS CATALYTIC SYSTEMS, *J. Phys. Chem. C* **2017**, *121*, 25847–25863 (cited on p. 9).
- [138] J. Müller, THEORETICAL INVESTIGATIONS OF COVALENT MECHANOCHEMISTRY, *PhD Thesis*, Christian-Albrechts-University, Kiel, **2017** (cited on pp. 9, 40, 71, 95).
- [139] W. Sun, Y.-X. Yuan, OPTIMIZATION THEORY AND METHODS: NONLINEAR PROGRAMMING, *Vol. 1, Vol. 1*, Springer Science & Business Media, New York, **2006** (cited on pp. 13, 15, 16, 18).
- [140] L. T. Biegler, I. E. Grossmann, RETROSPECTIVE ON OPTIMIZATION, *Comput. Chem. Eng* **2004**, *28*, 1169–1192 (cited on p. 13).
- [141] F. Jensen, INTRODUCTION TO COMPUTATIONAL CHEMISTRY, 2nd edition, Wiley, West Sussex, **2007** (cited on pp. 15, 20, 30, 31, 39, 40, 166).
- [142] H. B. Schlegel, GEOMETRY OPTIMIZATION, *WIREs Comput. Mol. Sci.* **2011**, *1*, 790–809 (cited on pp. 15, 51, 53).
- [143] A. R. Conn, K. Scheinberg, L. N. Vicente, INTRODUCTION TO DERIVATIVE-FREE OPTIMIZATION, *Society for Industrial and Applied Mathematics*, **2009** (cited on p. 16).
- [144] M. J. Powell, THE BOBYQA ALGORITHM FOR BOUND CONSTRAINED OPTIMIZATION WITHOUT DERIVATIVES, *Cambridge NA Report NA2009/06* **2009** (cited on pp. 16, 74).
- [145] M. J. D. Powell, DEVELOPMENTS OF NEWUOA FOR MINIMIZATION WITHOUT DERIVATIVES, *IMA J. Numer. Anal.* **2008**, *28*, 649–664 (cited on pp. 16, 74).
- [146] U. C. Singh, P. A. Kollman, AN APPROACH TO COMPUTING ELECTROSTATIC CHARGES FOR MOLECULES, *J. Comput. Chem.* **1984**, *5*, 129–145 (cited on p. 19).
- [147] S. R. Cox, D. E. Williams, REPRESENTATION OF THE MOLECULAR ELECTROSTATIC POTENTIAL BY A NET ATOMIC CHARGE MODEL, *J. Comput. Chem.* **1981**, *2*, 304–323 (cited on p. 19).
- [148] L. E. Chirlian, M. M. Francl, ATOMIC CHARGES DERIVED FROM ELECTROSTATIC POTENTIALS: A DETAILED STUDY, *J. Comput. Chem.* **1987**, *8*, 894–905 (cited on p. 19).
- [149] T. R. Stouch, D. E. Williams, CONFORMATIONAL DEPENDENCE OF ELECTROSTATIC POTENTIAL-DERIVED CHARGES: STUDIES OF THE FITTING PROCEDURE, *J. Comput. Chem.* **1993**, *14*, 858–866 (cited on p. 20).
- [150] M. M. Francl, C. Carey, L. E. Chirlian, D. M. Gange, CHARGES FIT TO ELECTROSTATIC POTENTIALS. II. CAN ATOMIC CHARGES BE UNAMBIGUOUSLY FIT TO ELECTROSTATIC POTENTIALS?, *J. Comput. Chem.* **1996**, *17*, 367–383 (cited on p. 20).
- [151] E. Sigfridsson, U. Ryde, COMPARISON OF METHODS FOR DERIVING ATOMIC CHARGES FROM THE ELECTROSTATIC POTENTIAL AND MOMENTS, *J. Comput. Chem.* **1998**, *19*, 377–395 (cited on p. 20).
- [152] C. J. Cramer, ESSENTIALS OF COMPUTATIONAL CHEMISTRY: THEORIES AND MODELS, 2nd edition, Wiley, West Sussex, **2004** (cited on pp. 20, 30, 31, 34).
- [153] W. D. Cornell, P. Cieplak, C. I. Bayly, P. A. Kollman, APPLICATION OF RESP CHARGES TO CALCULATE CONFORMATIONAL ENERGIES, HYDROGEN BOND ENERGIES, AND FREE ENERGIES OF SOLVATION, *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631 (cited on pp. 20, 166).
- [154] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, A WELL-BEHAVED ELECTROSTATIC POTENTIAL BASED METHOD USING CHARGE RESTRAINTS FOR DERIVING ATOMIC CHARGES: THE RESP MODEL, *J. Phys. Chem.* **1993**, *97*, 10269–10280 (cited on pp. 20, 166).
- [155] T. Weise, R. Chiong, K. Táng, EVOLUTIONARY OPTIMIZATION: PITFALLS AND BOOBY TRAPS, *J. Comp. Sci. Technol.* **2012**, *27*, 907–936 (cited on pp. 21, 22, 147).

- [156] S. Luke, ESSENTIALS OF METAHEURISTICS, 2nd edition (online version 2.2), Lulu, 2013 (cited on pp. 21, 24, 25, 27–29, 73, 78, 147, 156).
- [157] R. Bellman, DYNAMIC PROGRAMMING, Princeton University Press, 1957 (cited on p. 21).
- [158] H. R. Larsson, EFFICIENT APPROACHES TO MULTIDIMENSIONAL QUANTUM DYNAMICS: DYNAMICAL PRUNING IN PHASE, POSITION AND CONFIGURATION SPACE, PhD Thesis, Christian-Albrechts University, Kiel, 2018 (cited on p. 21).
- [159] M. Rupp, KERNEL METHODS FOR VIRTUAL SCREENING, PhD Thesis, University of Frankfurt, Frankfurt, 2009 (cited on pp. 21, 67).
- [160] F. H. Stillinger, T. A. Weber, HIDDEN STRUCTURE IN LIQUIDS, *Phys. Rev. A* **1982**, *25*, 978–989 (cited on p. 21).
- [161] L. T. Wille, J. Vennik, COMPUTATIONAL COMPLEXITY OF THE GROUND-STATE DETERMINATION OF ATOMIC CLUSTERS, *J. Phys. A: Math. Gen.* **1985**, *18*, L419 (cited on pp. 21, 23).
- [162] G. W. Greenwood, REVISITING THE COMPLEXITY OF FINDING GLOBALLY MINIMUM ENERGY CONFIGURATIONS IN ATOMIC CLUSTERS, *Z. Phys. Chem.* **1999**, *211*, 105–114 (cited on pp. 21, 23).
- [163] J. P. K. Doye, D. J. Wales, SADDLE POINTS AND DYNAMICS OF LENNARD-JONES CLUSTERS, SOLIDS, AND SUPERCOOLED LIQUIDS, *J. Chem. Phys.* **2002**, *116*, 3777–3788 (cited on p. 21).
- [164] D. J. Wales, ENERGY LANDSCAPES: APPLICATIONS TO CLUSTERS, BIOMOLECULES AND GLASSES, Cambridge University Press, Cambridge, 2003 (cited on pp. 21, 117).
- [165] D. J. Wales, EXPLORING ENERGY LANDSCAPES, *Annu. Rev. Phys. Chem.* **2018**, *69*, 401–425 (cited on pp. 21, 117).
- [166] B. Hartke, GLOBAL OPTIMIZATION, *WIREs Comput. Mol. Sci.* **2011**, *1*, 879–887 (cited on p. 23).
- [167] B. Hartke, GLOBAL CLUSTER GEOMETRY OPTIMIZATION BY A PHENOTYPE ALGORITHM WITH NICHES: LOCATION OF ELUSIVE MINIMA, AND LOW-ORDER SCALING WITH CLUSTER SIZE, *J. Comput. Chem.* **1999**, *20*, 1752–1759 (cited on pp. 23, 28, 74, 78, 117).
- [168] A. Neumaier, O. Shcherbina, W. Huyer, T. Vinkó, A COMPARISON OF COMPLETE GLOBAL OPTIMIZATION SOLVERS, *Math. Program.* **2005**, *103*, 335–356 (cited on p. 23).
- [169] C. A. Floudas, C. E. Gounaris, A REVIEW OF RECENT ADVANCES IN GLOBAL OPTIMIZATION, *J. Glob. Optim.* **2009**, *45*, 3 (cited on p. 23).
- [170] F. Glover, FUTURE PATHS FOR INTEGER PROGRAMMING AND LINKS TO ARTIFICIAL INTELLIGENCE, *Comput. Oper. Res.* **1986**, *13*, 533–549 (cited on p. 24).
- [171] L. Y. P. Luk, J. J. Ruiz-Pernía, W. M. Dawson, M. Roca, E. J. Loveridge, D. R. Glowacki, J. N. Harvey, A. J. Mulholland, I. Tuñón, V. Moliner, R. K. Allemann, UNRAVELING THE ROLE OF PROTEIN DYNAMICS IN DIHYDROFOLATE REDUCTASE CATALYSIS, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 16344–16349 (cited on p. 25).
- [172] D. Wolf, O. Buyevskaya, M. Baerns, AN EVOLUTIONARY APPROACH IN THE COMBINATORIAL SELECTION AND OPTIMIZATION OF CATALYTIC MATERIALS, *Appl. Catal. A* **2000**, *200*, 63–77 (cited on p. 25).
- [173] Z. Li, H. A. Scheraga, MONTE CARLO-MINIMIZATION APPROACH TO THE MULTIPLE-MINIMA PROBLEM IN PROTEIN FOLDING, *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6611–6615 (cited on p. 26).
- [174] D. J. Wales, J. P. K. Doye, GLOBAL OPTIMIZATION BY BASIN-HOPPING AND THE LOWEST ENERGY STRUCTURES OF LENNARD-JONES CLUSTERS CONTAINING UP TO 110 ATOMS, *J. Phys. Chem. A* **1997**, *101*, 5111–5116 (cited on pp. 26, 117).

- [175] D. J. Wales, H. A. Scheraga, GLOBAL OPTIMIZATION OF CLUSTERS, CRYSTALS, AND BIOMOLECULES, *Science* **1999**, *285*, 1368–1372 (cited on p. 26).
- [176] Weise, Why Research in Computational Intelligence Should Be Less Nature-Inspired. <http://iao.hfuu.edu.cn/blogs/21-why-research-in-computational-intelligence-should-be-less-nature-inspired> (visited on 11/29/2017) (cited on pp. 26–28).
- [177] O. N. Carstensen, J. M. Dieterich, B. Hartke, DESIGN OF OPTIMALLY SWITCHABLE MOLECULES BY GENETIC ALGORITHMS, *Phys. Chem. Chem. Phys.* **2011**, *13*, 2903–2910 (cited on pp. 28, 70, 229).
- [178] C. Rupakheti, A. Virshup, W. Yang, D. N. Beratan, STRATEGY TO DISCOVER DIVERSE OPTIMAL MOLECULES IN THE SMALL MOLECULE UNIVERSE, *J. Chem. Inf. Model.* **2015**, *55*, 529–537 (cited on p. 28).
- [179] M. d’Avezac, A. Zunger, IDENTIFYING THE MINIMUM-ENERGY ATOMIC CONFIGURATION ON A LATTICE: LAMARCKIAN TWIST ON DARWINIAN EVOLUTION, *Phys. Rev. B* **2008**, *78*, 064102 (cited on p. 28).
- [180] D. M. Daeven, N. Tit, J. R. Morris, K. M. Ho, STRUCTURAL OPTIMIZATION OF LENNARD-JONES CLUSTERS BY A GENETIC ALGORITHM, *Chem. Phys. Lett.* **1996**, *256*, 195–200 (cited on pp. 28, 78, 117).
- [181] W. J. Pullan, GENETIC OPERATORS FOR THE ATOMIC CLUSTER PROBLEM, *Comput. Phys. Commun.* **1997**, *107*, 137–148 (cited on pp. 28, 78).
- [182] F. C. Chuang, C. V. Ciobanu, V. B. Shenoy, C. Z. Wang, K. M. Ho, FINDING THE RECONSTRUCTIONS OF SEMICONDUCTOR SURFACES VIA A GENETIC ALGORITHM, *Surf. Sci.* **2004**, *573*, L375–L381 (cited on p. 28).
- [183] J. Holland, ADAPTATION IN NATURAL AND ARTIFICIAL SYSTEMS: AN INTRODUCTORY ANALYSIS WITH APPLICATIONS TO BIOLOGY, CONTROL, AND ARTIFICIAL INTELLIGENCE, University of Michigan Press, **1975** (cited on p. 28).
- [184] A. Szabo, N. S. Ostlund, MODERN QUANTUM CHEMISTRY: INTRODUCTION TO ADVANCED ELECTRONIC STRUCTURE THEORY, Dover Publications Inc., Mineola, **1996** (cited on pp. 30, 31).
- [185] P. W. Atkins, R. S. Friedman, MOLECULAR QUANTUM MECHANICS, 5th edition, Oxford University Press, Oxford, **2010** (cited on pp. 30, 31).
- [186] L. Piela, IDEAS OF QUANTUM CHEMISTRY, 2nd edition, Elsevier, Amsterdam, **2014** (cited on pp. 30, 31, 43–45, 56).
- [187] I. N. Levine, QUANTUM CHEMISTRY, 7th edition, Pearson, Boston, **2014** (cited on pp. 30, 31).
- [188] M. Born, R. Oppenheimer, ZUR QUANTENTHEORIE DER MOLEKELN, *Ann. Phys.* **1927**, *389*, 457–484 (cited on p. 30).
- [189] A. V. Akimov, O. V. Prezhdo, LARGE-SCALE COMPUTATIONS IN CHEMISTRY: A BIRD’S EYE VIEW OF A VIBRANT FIELD, *Chem. Rev.* **2015**, *115*, 5797–5890 (cited on p. 31).
- [190] P. a. M. Dirac, A NEW NOTATION FOR QUANTUM MECHANICS, *Math. Proc. Camb. Philos. Soc.* **1939**, *35*, 416–418 (cited on p. 32).
- [191] J. Binney, THE PHYSICS OF QUANTUM MECHANICS, Oxford University Press, Oxford, **2013** (cited on p. 32).
- [192] D. P. Tew, W. Klopper, T. Helgaker, ELECTRON CORRELATION: THE MANY-BODY PROBLEM AT THE HEART OF CHEMISTRY, *J. Comput. Chem.* **2007**, *28*, 1307–1320 (cited on p. 33).
- [193] T. Helgaker, J. Olsen, P. Jørgensen, MOLECULAR ELECTRONIC-STRUCTURE THEORY, 1st edition, John Wiley & Sons, Ltd, Chichester, **2013** (cited on p. 34).

- [194] J. B. Schönborn, DYNAMICS OF PHOTOINDUCED SWITCHING PROCESSES, *PhD Thesis*, Christian-Albrechts-University, Kiel, **2013** (cited on p. 34).
- [195] N. O. Cartensen, SIMULATING THE PHOTODYNAMICS OF AZOBENZENE DERIVATIVES USING SURFACE-HOPPING DYNAMICS, *PhD Thesis*, Christian-Albrechts-University, Kiel, **2014** (cited on pp. 34, 51).
- [196] T. Raeker, FULL-DIMENSIONAL PHOTODYNAMICS SIMULATIONS – FROM PHOTOISOMERIZATIONS TO EXCITED-STATE PROTON TRANSFER SYSTEMS, *PhD Thesis*, Christian-Albrechts-University, Kiel, **2018** (cited on pp. 34, 37, 51).
- [197] P. Hohenberg, W. Kohn, INHOMOGENEOUS ELECTRON GAS, *Phys. Rev.* **1964**, *136*, B864–B871 (cited on p. 34).
- [198] W. Kohn, L. J. Sham, SELF-CONSISTENT EQUATIONS INCLUDING EXCHANGE AND CORRELATION EFFECTS, *Phys. Rev.* **1965**, *140*, A1133–A1138 (cited on p. 35).
- [199] L. Goerigk, A. Hansen, C. A. Bauer, S. Ehrlich, A. Najibi, S. Grimme, A LOOK AT THE DENSITY FUNCTIONAL THEORY ZOO WITH THE ADVANCED GMTKN55 DATABASE FOR GENERAL MAIN GROUP THERMOCHEMISTRY, KINETICS AND NONCOVALENT INTERACTIONS, *Phys. Chem. Chem. Phys.* **2017** (cited on p. 36).
- [200] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, K. A. Lyssenko, DENSITY FUNCTIONAL THEORY IS STRAYING FROM THE PATH TOWARD THE EXACT FUNCTIONAL, *Science* **2017**, *355*, 49–52 (cited on p. 36).
- [201] C. Adamo, V. Barone, TOWARD RELIABLE DENSITY FUNCTIONAL METHODS WITHOUT ADJUSTABLE PARAMETERS: THE PBE0 MODEL, *J. Chem. Phys.* **1999**, *110*, 6158–6170 (cited on pp. 36, 158).
- [202] W. Thiel, SEMIEMPIRICAL QUANTUM–CHEMICAL METHODS, *WIREs Comput. Mol. Sci.* **2014**, *4*, 145–157 (cited on pp. 36, 37, 39).
- [203] J. A. Pople, D. L. Beveridge, P. A. Dobosh, APPROXIMATE SELF-CONSISTENT MOLECULAR-ORBITAL THEORY. V. INTERMEDIATE NEGLECT OF DIFFERENTIAL OVERLAP, *J. Chem. Phys.* **1967**, *47*, 2026–2033 (cited on p. 36).
- [204] J. A. Pople, D. P. Santry, G. A. Segal, APPROXIMATE SELF-CONSISTENT MOLECULAR ORBITAL THEORY. I. INVARIANT PROCEDURES, *J. Chem. Phys.* **1965**, *43*, S129–S135 (cited on p. 36).
- [205] J. A. Pople, D. L. Beveridge, APPROXIMATE MOLECULAR ORBITAL THEORY, McGraw-Hill, New York, **1970** (cited on p. 37).
- [206] J. J. P. Stewart, OPTIMIZATION OF PARAMETERS FOR SEMIEMPIRICAL METHODS VI: MORE MODIFICATIONS TO THE NDDO APPROXIMATIONS AND RE-OPTIMIZATION OF PARAMETERS, *J. Mol. Model.* **2013**, *19*, 1–32 (cited on pp. 37, 158, 182).
- [207] M. J. S. Dewar, W. Thiel, GROUND STATES OF MOLECULES. 38. THE MNDO METHOD. APPROXIMATIONS AND PARAMETERS, *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907 (cited on p. 37).
- [208] M. J. S. Dewar, W. Thiel, GROUND STATES OF MOLECULES. 39. MNDO RESULTS FOR MOLECULES CONTAINING HYDROGEN, CARBON, NITROGEN, AND OXYGEN, *J. Am. Chem. Soc.* **1977**, *99*, 4907–4917 (cited on p. 37).
- [209] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, DEVELOPMENT AND USE OF QUANTUM MECHANICAL MOLECULAR MODELS. 76. AM1: A NEW GENERAL PURPOSE QUANTUM MECHANICAL MOLECULAR MODEL, *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909 (cited on p. 37).
- [210] J. J. P. Stewart, OPTIMIZATION OF PARAMETERS FOR SEMIEMPIRICAL METHODS I. METHOD, *J. Comput. Chem.* **1989**, *10*, 209–220 (cited on p. 37).
- [211] J. J. P. Stewart, OPTIMIZATION OF PARAMETERS FOR SEMIEMPIRICAL METHODS II. APPLICATIONS, *J. Comput. Chem.* **1989**, *10*, 221–264 (cited on p. 37).

- [212] W. Thiel, A. A. Voityuk, EXTENSION OF THE MNDO FORMALISM TO D ORBITALS: INTEGRAL APPROXIMATIONS AND PRELIMINARY NUMERICAL RESULTS, *Theor. Chim. Acta* **1992**, *81*, 391–404 (cited on p. 37).
- [213] W. Thiel, A. A. Voityuk, EXTENSION OF MNDO TO D ORBITALS: PARAMETERS AND RESULTS FOR THE SECOND-ROW ELEMENTS AND FOR THE ZINC GROUP, *J. Phys. Chem.* **1996**, *100*, 616–626 (cited on p. 37).
- [214] J. J. P. Stewart, OPTIMIZATION OF PARAMETERS FOR SEMIEMPIRICAL METHODS V: MODIFICATION OF NDDO APPROXIMATIONS AND APPLICATION TO 70 ELEMENTS, *J. Mol. Model.* **2007**, *13*, 1173–1213 (cited on pp. 37, 60).
- [215] G. B. Rocha, R. O. Freire, A. M. Simas, J. J. P. Stewart, RM1: A REPARAMETERIZATION OF AM1 FOR H, C, N, O, P, S, F, CL, BR, AND I, *J. Comput. Chem.* **2006**, *27*, 1101–1111 (cited on p. 37).
- [216] T. Husch, A. C. Vaucher, M. Reiher, SEMIEMPIRICAL MOLECULAR ORBITAL MODELS BASED ON THE NEGLECT OF DIATOMIC DIFFERENTIAL OVERLAP APPROXIMATION, *Int. J. Quantum Chem.* **2018**, *118*, e25799 (cited on pp. 37, 38).
- [217] M. Kolb, W. Thiel, BEYOND THE MNDO MODEL: METHODOLOGICAL CONSIDERATIONS AND NUMERICAL RESULTS, *J. Comput. Chem.* **1993**, *14*, 775–789 (cited on pp. 37, 38).
- [218] W. Weber, W. Thiel, ORTHOGONALIZATION CORRECTIONS FOR SEMIEMPIRICAL METHODS, *Theor. Chem. Acc.* **2000**, *103*, 495–506 (cited on pp. 37, 38).
- [219] P. O. Dral, X. Wu, W. Thiel, SEMIEMPIRICAL QUANTUM-CHEMICAL METHODS WITH ORTHOGONALIZATION AND DISPERSION CORRECTIONS, *J. Chem. Theory Comput.* **2019**, *15*, 1743–1760 (cited on p. 37).
- [220] A. S. Christensen, T. Kubař, Q. Cui, M. Elstner, SEMIEMPIRICAL QUANTUM MECHANICAL METHODS FOR NONCOVALENT INTERACTIONS FOR CHEMICAL AND BIOCHEMICAL APPLICATIONS, *Chem. Rev.* **2016**, *116*, 5301–5337 (cited on pp. 37, 38).
- [221] P. O. Dral, X. Wu, L. Spörkel, A. Kosłowski, W. Weber, R. Steiger, M. Scholten, W. Thiel, SEMIEMPIRICAL QUANTUM-CHEMICAL ORTHOGONALIZATION-CORRECTED METHODS: THEORY, IMPLEMENTATION, AND PARAMETERS, *J. Chem. Theory Comput.* **2016**, *12*, 1082–1096 (cited on p. 38).
- [222] P. O. Dral, X. Wu, L. Spörkel, A. Kosłowski, W. Thiel, SEMIEMPIRICAL QUANTUM-CHEMICAL ORTHOGONALIZATION-CORRECTED METHODS: BENCHMARKS FOR GROUND-STATE PROPERTIES, *J. Chem. Theory Comput.* **2016**, *12*, 1097–1120 (cited on p. 38).
- [223] D. Tuna, Y. Lu, A. Kosłowski, W. Thiel, SEMIEMPIRICAL QUANTUM-CHEMICAL ORTHOGONALIZATION-CORRECTED METHODS: BENCHMARKS OF ELECTRONICALLY EXCITED STATES, *J. Chem. Theory Comput.* **2016**, *12*, 4400–4422 (cited on p. 38).
- [224] X. Wu, P. O. Dral, A. Kosłowski, W. Thiel, BIG DATA ANALYSIS OF AB INITIO MOLECULAR INTEGRALS IN THE NEGLECT OF DIATOMIC DIFFERENTIAL OVERLAP APPROXIMATION, *J. Comput. Chem.* **2019**, *40*, 638–649 (cited on p. 38).
- [225] T. Husch, M. Reiher, COMPREHENSIVE ANALYSIS OF THE NEGLECT OF DIATOMIC DIFFERENTIAL OVERLAP APPROXIMATION, *J. Chem. Theory Comput.* **2018**, *14*, 5169–5179 (cited on p. 38).
- [226] J. J. P. Stewart, MOPAC2016, STEWART COMPUTATIONAL CHEMISTRY, COLORADO SPRINGS, CO, USA, **2016**, <http://openmopac.net/> (visited on 03/11/2019) (cited on pp. 38, 51, 76, 182).
- [227] S. Grimme, A GENERAL QUANTUM MECHANICALLY DERIVED FORCE FIELD (QMDF) FOR MOLECULES AND CONDENSED PHASE SIMULATIONS, *J. Chem. Theory Comput.* **2014**, *10*, 4497–4514 (cited on pp. 38, 39, 41, 72).

- [228] S. Grimme, C. Bannwarth, P. Shushkov, A ROBUST AND ACCURATE TIGHT-BINDING QUANTUM CHEMICAL METHOD FOR STRUCTURES, VIBRATIONAL FREQUENCIES, AND NONCOVALENT INTERACTIONS OF LARGE MOLECULAR SYSTEMS PARAMETRIZED FOR ALL SPD-BLOCK ELEMENTS ($Z = 1-86$), *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009 (cited on pp. 38, 76).
- [229] C. Bannwarth, S. Ehlert, S. Grimme, GFN2-xTB—AN ACCURATE AND BROADLY PARAMETRIZED SELF-CONSISTENT TIGHT-BINDING QUANTUM CHEMICAL METHOD WITH MULTIPOLE ELECTROSTATICS AND DENSITY-DEPENDENT DISPERSION CONTRIBUTIONS, *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671 (cited on pp. 38, 76).
- [230] P. Koskinen, V. Mäkinen, DENSITY-FUNCTIONAL TIGHT-BINDING FOR BEGINNERS, *Comput. Mater. Sci.* **2009**, *47*, 237–253 (cited on p. 38).
- [231] M. Gaus, Q. Cui, M. Elstner, DFTB3: EXTENSION OF THE SELF-CONSISTENT-CHARGE DENSITY-FUNCTIONAL TIGHT-BINDING METHOD (SCC-DFTB), *J. Chem. Theory Comput.* **2011**, *7*, 931–948 (cited on p. 38).
- [232] S. Grimme, EXPLORATION OF CHEMICAL COMPOUND, CONFORMER, AND REACTION SPACE WITH META-DYNAMICS SIMULATIONS BASED ON TIGHT-BINDING QUANTUM CHEMICAL CALCULATIONS, *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862 (cited on p. 38).
- [233] L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande, T. J. Martínez, DISCOVERING CHEMISTRY WITH AN AB INITIO NANOREACTOR, *Nat. Chem.* **2014**, *6*, 1044–1048 (cited on pp. 38, 231).
- [234] M. Bursch, H. Neugebauer, S. Grimme, STRUCTURE OPTIMISATION OF LARGE TRANSITION METAL COMPLEXES WITH EXTENDED TIGHT-BINDING METHODS, *Angew. Chem.* **2019** (cited on p. 38).
- [235] K. Farah, F. Müller-Plathe, M. C. Böhm, CLASSICAL REACTIVE MOLECULAR DYNAMICS IMPLEMENTATIONS: STATE OF THE ART, *ChemPhysChem* **2012**, *13*, 1127–1151 (cited on p. 39).
- [236] A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, REAXFF: A REACTIVE FORCE FIELD FOR HYDROCARBONS, *J. Phys. Chem. A* **2001**, *105*, 9396–9409 (cited on pp. 39, 95).
- [237] A. C. T. van Duin, A. Strachan, S. Stewman, Q. Zhang, X. Xu, W. A. Goddard, REAXFF_{SiO} REACTIVE FORCE FIELD FOR SILICON AND SILICON OXIDE SYSTEMS, *J. Phys. Chem. A* **2003**, *107*, 3803–3811 (cited on pp. 39, 95).
- [238] A. Warshel, R. M. Weiss, AN EMPIRICAL VALENCE BOND APPROACH FOR COMPARING REACTIONS IN SOLUTIONS AND IN ENZYMES, *J. Am. Chem. Soc.* **1980**, *102*, 6218–6226 (cited on pp. 39, 40).
- [239] S. C. L. Kamerlin, A. Warshel, THE EMPIRICAL VALENCE BOND MODEL: THEORY AND APPLICATIONS, *WIREs Comput. Mol. Sci.* **2011**, *1*, 30–45 (cited on pp. 39, 40).
- [240] M. Dittner, *NEUE IMPLEMENTATION GLOBALER PARAMETEROPTIMIERUNG EINES REAKTIVEN KRAFTFELDES*, Master's Thesis, Christian-Albrechts-University, Kiel, **2014** (cited on pp. 39, 71, 74, 78, 95, 96, 110).
- [241] W. J. Mortier, S. K. Ghosh, S. Shankar, ELECTRONEGATIVITY-EQUALIZATION METHOD FOR THE CALCULATION OF ATOMIC CHARGES IN MOLECULES, *J. Am. Chem. Soc.* **1986**, *108*, 4315–4320 (cited on p. 39).
- [242] A. K. Rappe, W. A. Goddard, CHARGE EQUILIBRATION FOR MOLECULAR DYNAMICS SIMULATIONS, *J. Phys. Chem.* **1991**, *95*, 3358–3363 (cited on p. 39).
- [243] T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, A. C. T. van Duin, THE REAXFF REACTIVE FORCE-FIELD: DEVELOPMENT, APPLICATIONS AND FUTURE DIRECTIONS, *Npj Comput. Mater.* **2016**, *2*, 15011 (cited on pp. 39, 40).

- [244] M. Dittner, J. Müller, H. M. Aktulga, B. Hartke, EFFICIENT GLOBAL OPTIMIZATION OF REACTIVE FORCE-FIELD PARAMETERS, *J. Comput. Chem.* **2015**, *36*, 1550–1561 (cited on pp. 40, 71, 72, 96, 109, 287).
- [245] H. R. Larsson, A. C. T. van Duin, B. Hartke, GLOBAL OPTIMIZATION OF PARAMETERS IN THE REACTIVE FORCE FIELD REAXFF FOR SiOH, *J. Comput. Chem.* **2013**, *34*, 2178–2189 (cited on p. 40).
- [246] H. R. Larsson, B. Hartke, FITTING REACTIVE FORCE FIELDS USING GENETIC ALGORITHMS, *Comput. Meth. Mater. Sci.* **2013**, *13*, 120–126 (cited on p. 40).
- [247] J. Müller, B. Hartke, REAXFF REACTIVE FORCE FIELD FOR DISULFIDE MECHANOCHEMISTRY, FITTED TO MULTIREFERENCE AB INITIO DATA, *J. Chem. Theory Comput.* **2016**, *12*, 3913–3925 (cited on pp. 40, 71, 95).
- [248] L.-P. Wang, T. J. Martinez, V. S. Pande, BUILDING FORCE FIELDS: AN AUTOMATIC, SYSTEMATIC, AND REPRODUCIBLE APPROACH, *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891 (cited on p. 40).
- [249] A. Jaramillo-Botero, S. Naserifar, W. A. Goddard, GENERAL MULTIOBJECTIVE FORCE FIELD OPTIMIZATION FRAMEWORK, WITH APPLICATION TO REACTIVE FORCE FIELDS FOR SILICON CARBIDE, *J. Chem. Theory Comput.* **2014**, *10*, 1426–1439 (cited on p. 40).
- [250] J. P. Larentzos, B. M. Rice, E. F. C. Byrd, N. S. Weingarten, J. V. Lill, PARAMETERIZING COMPLEX REACTIVE FORCE FIELDS USING MULTIPLE OBJECTIVE EVOLUTIONARY STRATEGIES (MOES). PART 1: REAXFF MODELS FOR CYCLOTTRIMETHYLENE TRINITRAMINE (RDX) AND 1,1-DIAMINO-2,2-DINITROETHENE (FOX-7), *J. Chem. Theory Comput.* **2015**, *11*, 381–391 (cited on p. 40).
- [251] B. M. Rice, J. P. Larentzos, E. F. C. Byrd, N. S. Weingarten, PARAMETERIZING COMPLEX REACTIVE FORCE FIELDS USING MULTIPLE OBJECTIVE EVOLUTIONARY STRATEGIES (MOES): PART 2: TRANSFERABILITY OF REAXFF MODELS TO C–H–N–O ENERGETIC MATERIALS, *J. Chem. Theory Comput.* **2015**, *11*, 392–405 (cited on p. 40).
- [252] A. C. T. van Duin, J. M. A. Baas, B. van de Graaf, DELFT MOLECULAR MECHANICS: A NEW APPROACH TO HYDROCARBON FORCE FIELDS. INCLUSION OF A GEOMETRY-DEPENDENT CHARGE CALCULATION, *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 2881–2895 (cited on p. 40).
- [253] K. Chenoweth, A. C. T. van Duin, W. A. Goddard, REAXFF REACTIVE FORCE FIELD FOR MOLECULAR DYNAMICS SIMULATIONS OF HYDROCARBON OXIDATION, *J. Phys. Chem. A* **2008**, *112*, 1040–1053 (cited on p. 40).
- [254] M. F. Russo, A. C. van Duin, ATOMISTIC-SCALE SIMULATIONS OF CHEMICAL REACTIONS: BRIDGING FROM QUANTUM CHEMISTRY TO ENGINEERING, *Nucl. Instr. Meth. Phys. Res. B* **2011**, *269*, 1549–1554 (cited on p. 40).
- [255] T. Liang, Y. K. Shin, Y.-T. Cheng, D. E. Yilmaz, K. G. Vishnu, O. Veners, C. Zou, S. R. Phillpot, S. B. Sinnott, A. C. van Duin, REACTIVE POTENTIALS FOR ADVANCED ATOMISTIC SIMULATIONS, *Annu. Rev. Mater. Res.* **2013**, *43*, 109–129 (cited on p. 40).
- [256] Y. K. Shin, T.-R. Shan, T. Liang, M. J. Noordhoek, S. B. Sinnott, A. C. van Duin, S. R. Phillpot, VARIABLE CHARGE MANY-BODY INTERATOMIC POTENTIALS, *MRS Bull.* **2012**, *37*, 504–512 (cited on p. 40).
- [257] Frank, Per-Ola, TRANSITION STATES FROM EMPIRICAL FORCE FIELDS, *Theor. Chem. Acc.* **2003**, *109*, 1–7 (cited on p. 41).
- [258] J. Steffen, B. Hartke, CHEAP BUT ACCURATE CALCULATION OF CHEMICAL REACTION RATE CONSTANTS FROM AB INITIO DATA, VIA SYSTEM-SPECIFIC, BLACK-BOX FORCE FIELDS, *J. Chem. Phys.* **2017**, *147*, 161701 (cited on pp. 41, 42).

- [259] J. Steffen, A NEW CLASS OF REACTION PATH BASED POTENTIAL ENERGY SURFACES ENABLING ACCURATE BLACK BOX CHEMICAL RATE CONSTANT CALCULATIONS, *J. Chem. Phys.* **2019**, *150*, 154105 (cited on p. 41).
- [260] C. Bannwarth, J. Seibert, S. Grimme, ELECTRONIC CIRCULAR DICHROISM OF [16]HELICENE WITH SIMPLIFIED TD-DFT: BEYOND THE SINGLE STRUCTURE APPROACH, *Chirality* **2016**, *28*, 365–369 (cited on p. 41).
- [261] A. V. Marenich, S. V. Jerome, C. J. Cramer, D. G. Truhlar, CHARGE MODEL 5: AN EXTENSION OF HIRSHFELD POPULATION ANALYSIS FOR THE ACCURATE DESCRIPTION OF MOLECULAR INTERACTIONS IN GASEOUS AND CONDENSED PHASES, *J. Chem. Theory Comput.* **2012**, *8*, 527–541 (cited on p. 41).
- [262] I. Mayer, CHARGE, BOND ORDER AND VALENCE IN THE AB INITIO SCF THEORY, *Chem. Phys. Lett.* **1983**, *97*, 270–274 (cited on p. 41).
- [263] K. B. Wiberg, APPLICATION OF THE POPLE-SANTRY-SEGAL CNDO METHOD TO THE CYCLO-PROPYLCARBINYL AND CYCLOBUTYL CATION AND TO BICYCLOBUTANE, *Tetrahedron* **1968**, *24*, 1083–1096 (cited on p. 41).
- [264] B. Hartke, S. Grimme, REACTIVE FORCE FIELDS MADE SIMPLE, *Phys. Chem. Chem. Phys.* **2015**, *17*, 16715–16718 (cited on pp. 41, 72, 76, 90, 148).
- [265] J. Steffen, REAKTIVE KRAFTFELDER AUF BASIS QUANTENCHEMISCHER RECHNUNGEN: IMPLEMENTATION UND ANWENDUNGEN, Master Thesis, Christian-Albrechts-University, Kiel, **2015** (cited on p. 42).
- [266] C. A. Coulson, ELECTRICITY, 5th edition, Oliver and Boyd, Interscience-Publishers, New York, **1958** (cited on p. 42).
- [267] A. J. Stone, THE THEORY OF INTERMOLECULAR FORCES, Oxford University Press, Oxford, **1997** (cited on pp. 43, 44).
- [268] S. Shaik, R. Ramanan, D. Danovich, D. Mandal, STRUCTURE AND REACTIVITY/SELECTIVITY CONTROL BY ORIENTED-EXTERNAL ELECTRIC FIELDS, *Chem. Soc. Rev.* **2018**, *47*, 5125–5145 (cited on pp. 47, 177–179, 181, 208, 212, 227).
- [269] N. O. Carstensen, QM/MM SURFACE-HOPPING DYNAMICS OF A BRIDGED AZOBENZENE DERIVATIVE, *Phys. Chem. Chem. Phys.* **2013**, *15*, 15017 (cited on p. 47).
- [270] K. E. Ranaghan, L. Ridder, B. Szeferczyk, W. A. Sokalski, J. C. Hermann, A. J. Mulholland, TRANSITION STATE STABILIZATION AND SUBSTRATE STRAIN IN ENZYME CATALYSIS: AB INITIO QM/MM MODELLING OF THE CHORISMATE MUTASE REACTION, *Org. Biomol. Chem.* **2004**, *2*, 968–980 (cited on p. 47).
- [271] F. Claeysens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel, H.-J. Werner, HIGH-ACCURACY COMPUTATION OF REACTION BARRIERS IN ENZYMES, *Angew. Chem. Int. Ed.* **2006**, *45*, 6856–6859 (cited on p. 47).
- [272] S. Dapprich, I. Komáromi, K. S. Byun, K. Morokuma, M. J. Frisch, A NEW ONIOM IMPLEMENTATION IN GAUSSIAN98. PART I. THE CALCULATION OF ENERGIES, GRADIENTS, VIBRATIONAL FREQUENCIES AND ELECTRIC FIELD DERIVATIVES DEDICATED TO PROFESSOR KEIJI MOROKUMA IN CELEBRATION OF HIS 65TH BIRTHDAY.1, *J. Mol. Struct. THEOCHEM* **1999**, *461-462*, 1–21 (cited on p. 47).
- [273] A. Warshel, M. Levitt, THEORETICAL STUDIES OF ENZYMIC REACTIONS: DIELECTRIC, ELECTROSTATIC AND STERIC STABILIZATION OF THE CARBONIUM ION IN THE REACTION OF LYSOZYME, *J. Mol. Biol.* **1976**, *103*, 227–249 (cited on pp. 47, 50).
- [274] H. M. Senn, W. Thiel, QM/MM METHODS FOR BIOMOLECULAR SYSTEMS, *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229 (cited on pp. 47, 48, 50).

- [275] D. Bakowies, W. Thiel, SEMIEMPIRICAL TREATMENT OF ELECTROSTATIC POTENTIALS AND PARTIAL CHARGES IN COMBINED QUANTUM MECHANICAL AND MOLECULAR MECHANICAL APPROACHES, *J. Comput. Chem.* **1996**, *17*, 87–108 (cited on pp. 47, 50).
- [276] D. Bakowies, W. Thiel, HYBRID MODELS FOR COMBINED QUANTUM MECHANICAL AND MOLECULAR MECHANICAL APPROACHES, *J. Phys. Chem.* **1996**, *100*, 10580–10594 (cited on pp. 47, 50).
- [277] I. Antes, W. Thiel, ADJUSTED CONNECTION ATOMS FOR COMBINED QUANTUM MECHANICAL AND MOLECULAR MECHANICAL METHODS, *J. Phys. Chem. A* **1999**, *103*, 9290–9295 (cited on pp. 48, 50).
- [278] P. Politzer, J. S. Murray, THE FUNDAMENTAL NATURE AND ROLE OF THE ELECTROSTATIC POTENTIAL IN ATOMS AND MOLECULES, *Theor. Chem. Acc.* **2002**, *108*, 134–142 (cited on p. 49).
- [279] J. S. Murray, P. Politzer, THE ELECTROSTATIC POTENTIAL: AN OVERVIEW, *WIREs Comput. Mol. Sci.* **2011**, *1*, 153–163 (cited on p. 49).
- [280] M. A. Thompson, G. K. Schenter, EXCITED STATES OF THE BACTERIOCHLOROPHYLL B DIMER OF RHODOPSEUDOMONAS VIRIDIS: A QM/MM STUDY OF THE PHOTOSYNTHETIC REACTION CENTER THAT INCLUDES MM POLARIZATION, *J. Phys. Chem.* **1995**, *99*, 6374–6386 (cited on p. 50).
- [281] G. P. Ford, B. Wang, NEW APPROACH TO THE RAPID SEMIEMPIRICAL CALCULATION OF MOLECULAR ELECTROSTATIC POTENTIALS BASED ON THE AM1 WAVE FUNCTION: COMPARISON WITH AB INITIO HG/6-31G* RESULTS, *J. Comput. Chem.* **1993**, *14*, 1101–1111 (cited on p. 50).
- [282] P. L. Cummins, J. E. Gready, ATOMIC CHARGES DERIVED FROM SEMIEMPIRICAL ELECTROSTATIC POTENTIALS; AN INTERACTION ENERGY METHOD, *Chem. Phys. Lett.* **1990**, *174*, 355–360 (cited on p. 50).
- [283] M. J. Field, P. A. Bash, M. Karplus, A COMBINED QUANTUM MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR DYNAMICS SIMULATIONS, *J. Comput. Chem.* **1990**, *11*, 700–733 (cited on p. 50).
- [284] V. Luzhkov, A. Warshel, MICROSCOPIC MODELS FOR QUANTUM MECHANICAL CALCULATIONS OF CHEMICAL PROCESSES IN SOLUTIONS: LD/AMPAC AND SCAAS/AMPAC CALCULATIONS OF SOLVATION ENERGIES, *J. Comput. Chem.* **1992**, *13*, 199–213 (cited on p. 50).
- [285] N. V. Plotnikov, A. Warshel, EXPLORING, REFINING, AND VALIDATING THE PARADYNAMICS QM/MM SAMPLING, *J. Phys. Chem. B* **2012**, *116*, 10342–10356 (cited on p. 51).
- [286] A. Toniolo, C. Ciminelli, G. Granucci, T. Laino, M. Persico, QM/MM CONNECTION ATOMS FOR THE MULTISTATE TREATMENT OF ORGANIC AND BIOLOGICAL MOLECULES, *Theor. Chem. Acc.* **2004**, *111*, 270–279 (cited on p. 51).
- [287] W. Quapp, D. Heidrich, ANALYSIS OF THE CONCEPT OF MINIMUM ENERGY PATH ON THE POTENTIAL ENERGY SURFACE OF CHEMICALLY REACTING SYSTEMS, *Theor. Chim. Acta* **1984**, *66*, 245–260 (cited on p. 51).
- [288] K. Fukui, FORMULATION OF THE REACTION COORDINATE, *J. Phys. Chem.* **1970**, *74*, 4161–4163 (cited on p. 51).
- [289] K. Fukui, THE PATH OF CHEMICAL REACTIONS - THE IRC APPROACH, *Acc. Chem. Res.* **1981**, *14*, 363–368 (cited on p. 51).
- [290] C. Gonzalez, H. B. Schlegel, AN IMPROVED ALGORITHM FOR REACTION PATH FOLLOWING, *J. Chem. Phys.* **1989**, *90*, 2154–2161 (cited on p. 51).
- [291] S. Maeda, Y. Harabuchi, Y. Ono, T. Taketsugu, K. Morokuma, INTRINSIC REACTION COORDINATE: CALCULATION, BIFURCATION, AND AUTOMATED SEARCH, *Int. J. Quantum Chem.* **2015**, *115*, 258–269 (cited on p. 51).

- [292] O. T. Unke, S. Brickel, M. Meuwly, SAMPLING REACTIVE REGIONS IN PHASE SPACE BY FOLLOWING THE MINIMUM DYNAMIC PATH, *J. Chem. Phys.* **2019**, *150*, 074107 (cited on p. 51).
- [293] K. Müller, L. D. Brown, LOCATION OF SADDLE POINTS AND MINIMUM ENERGY PATHS BY A CONSTRAINED SIMPLEX OPTIMIZATION PROCEDURE, *Theoret. Chim. Acta* **1979**, *53*, 75–93 (cited on p. 52).
- [294] L. Maragliano, A. Fischer, E. Vanden-Eijnden, G. Ciccotti, STRING METHOD IN COLLECTIVE VARIABLES: MINIMUM FREE ENERGY PATHS AND ISOCOMMITTOR SURFACES, *J. Chem. Phys.* **2006**, *125*, 024106 (cited on p. 52).
- [295] G. Díaz Leines, J. Rogal, COMPARISON OF MINIMUM-ACTION AND STEEPEST-DESCENT PATHS IN GRADIENT SYSTEMS, *Phys. Rev. E* **2016**, *93*, 022307 (cited on p. 52).
- [296] W. E, E. Vanden-Eijnden, TRANSITION-PATH THEORY AND PATH-FINDING ALGORITHMS FOR THE STUDY OF RARE EVENTS, *Annu. Rev. Phys. Chem.* **2010**, *61*, 391–420 (cited on p. 52).
- [297] D. J. Wales, PERSPECTIVE: INSIGHT INTO REACTION COORDINATES AND DYNAMICS FROM THE POTENTIAL ENERGY LANDSCAPE, *J. Chem. Phys.* **2015**, *142*, 130901 (cited on p. 52).
- [298] A. L. Dewyer, A. J. Argüelles, P. M. Zimmerman, METHODS FOR EXPLORING REACTION SPACE IN MOLECULAR SYSTEMS, *WIREs Comput. Mol. Sci.* **2017**, *8*, e1354 (cited on pp. 53, 59, 231).
- [299] Y. Zeng, P. Xiao, G. Henkelman, UNIFICATION OF ALGORITHMS FOR MINIMUM MODE OPTIMIZATION, *J. Chem. Phys.* **2014**, *140*, 044115 (cited on p. 53).
- [300] H. Jónsson, G. Mills, K. W. Jacobsen, NUDGED ELASTIC BAND METHOD FOR FINDING MINIMUM ENERGY PATHS OF TRANSITIONS, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, World Scientific, **1998**, pp. 385–404 (cited on pp. 53, 54, 60, 212).
- [301] G. Henkelman, H. Jónsson, IMPROVED TANGENT ESTIMATE IN THE NUDGED ELASTIC BAND METHOD FOR FINDING MINIMUM ENERGY PATHS AND SADDLE POINTS, *J. Chem. Phys.* **2000**, *113*, 9978–9985 (cited on pp. 53, 54).
- [302] G. Henkelman, B. P. Uberuaga, H. Jónsson, A CLIMBING IMAGE NUDGED ELASTIC BAND METHOD FOR FINDING SADDLE POINTS AND MINIMUM ENERGY PATHS, *J. Chem. Phys.* **2000**, *113*, 9901–9904 (cited on p. 53).
- [303] W. E, W. Ren, E. Vanden-Eijnden, STRING METHOD FOR THE STUDY OF RARE EVENTS, *Phys. Rev. B* **2002**, *66*, 052301 (cited on p. 53).
- [304] W. E, W. Ren, E. Vanden-Eijnden, SIMPLIFIED AND IMPROVED STRING METHOD FOR COMPUTING THE MINIMUM ENERGY PATHS IN BARRIER-CROSSING EVENTS, *J. Chem. Phys.* **2007**, *126*, 164103 (cited on p. 53).
- [305] B. Peters, A. Heyden, A. T. Bell, A. Chakraborty, A GROWING STRING METHOD FOR DETERMINING TRANSITION STATES: COMPARISON TO THE NUDGED ELASTIC BAND AND STRING METHODS, *J. Chem. Phys.* **2004**, *120*, 7877–7886 (cited on p. 53).
- [306] W. Quapp, A GROWING STRING METHOD FOR THE REACTION PATHWAY DEFINED BY A NEWTON TRAJECTORY, *J. Chem. Phys.* **2005**, *122*, 174106 (cited on p. 53).
- [307] S. K. Burger, W. Yang, QUADRATIC STRING METHOD FOR DETERMINING THE MINIMUM-ENERGY PATH BASED ON MULTIOBJECTIVE OPTIMIZATION, *J. Chem. Phys.* **2006**, *124*, 054109 (cited on p. 53).
- [308] P. M. Zimmerman, SINGLE-ENDED TRANSITION STATE FINDING WITH THE GROWING STRING METHOD, *J. Comput. Chem.* **2015**, *36*, 601–611 (cited on p. 53).
- [309] D. Sheppard, G. Henkelman, PATHS TO WHICH THE NUDGED ELASTIC BAND CONVERGES, *J. Comput. Chem.* **2011**, *32*, 1769–1771 (cited on p. 53).
- [310] W. Quapp, J. M. Bofill, A COMMENT TO THE NUDGED ELASTIC BAND METHOD, *J. Comput. Chem.* **2010**, *31*, 2526–2531 (cited on p. 53).

- [311] W. Quapp, CHEMICAL REACTION PATHS AND CALCULUS OF VARIATIONS, *Theor. Chem. Acc.* **2008**, *121*, 227–237 (cited on p. 53).
- [312] P. M. Zimmerman, AUTOMATED DISCOVERY OF CHEMICALLY REASONABLE ELEMENTARY REACTION STEPS, *J. Comput. Chem.* **2013**, *34*, 1385–1392 (cited on pp. 53, 231).
- [313] W. M. C. Sameera, S. Maeda, K. Morokuma, COMPUTATIONAL CATALYSIS USING THE ARTIFICIAL FORCE INDUCED REACTION METHOD, *Acc. Chem. Res.* **2016**, *49*, 763–773 (cited on pp. 53, 231).
- [314] L. D. Jacobson, A. D. Bochevarov, M. A. Watson, T. F. Hughes, D. Rinaldo, S. Ehrlich, T. B. Steinbrecher, S. Vaitheeswaran, D. M. Philipp, M. D. Halls, R. A. Friesner, AUTOMATED TRANSITION STATE SEARCH AND ITS APPLICATION TO DIVERSE TYPES OF ORGANIC REACTIONS, *J. Chem. Theory Comput.* **2017**, 5780–5797 (cited on pp. 53, 231).
- [315] G. N. Simm, M. Reiher, CONTEXT-DRIVEN EXPLORATION OF COMPLEX CHEMICAL REACTION NETWORKS, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119 (cited on pp. 53, 231).
- [316] S. Habershon, AUTOMATED PREDICTION OF CATALYTIC MECHANISM AND RATE LAW USING GRAPH-BASED REACTION PATH SAMPLING, *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798 (cited on pp. 53, 231).
- [317] H. C. Herbol, J. Stevenson, P. Clancy, COMPUTATIONAL IMPLEMENTATION OF NUDGED ELASTIC BAND, RIGID ROTATION, AND CORRESPONDING FORCE OPTIMIZATION, *J. Chem. Theory Comput.* **2017**, *13*, 3250–3259 (cited on p. 53).
- [318] S. Bahn, K. Jacobsen, AN OBJECT-ORIENTED SCRIPTING INTERFACE TO A LEGACY ELECTRONIC STRUCTURE CODE, *Comput. Sci. Eng.* **2002**, *4*, 56–66 (cited on pp. 53, 56).
- [319] E. L. Kolsbjerg, M. N. Groves, B. Hammer, AN AUTOMATED NUDGED ELASTIC BAND METHOD, *J. Chem. Phys.* **2016**, *145*, 094107 (cited on pp. 54, 57–59).
- [320] E. L. Kolsbjerg, M. N. Groves, B. Hammer, ERRATUM: “AN AUTOMATED NUDGED ELASTIC BAND METHOD” [J. CHEM. PHYS. 145, 094107 (2016)], *J. Chem. Phys.* **2018**, *148*, 029903 (cited on pp. 54, 57, 58).
- [321] D. Sheppard, R. Terrell, G. Henkelman, OPTIMIZATION METHODS FOR FINDING MINIMUM ENERGY PATHS, *J. Chem. Phys.* **2008**, *128*, 134106 (cited on pp. 55, 59, 61).
- [322] J.-W. Chu, B. L. Trout, B. R. Brooks, A SUPER-LINEAR MINIMIZATION SCHEME FOR THE NUDGED ELASTIC BAND METHOD, *J. Chem. Phys.* **2003**, *119*, 12708–12717 (cited on pp. 55, 61).
- [323] M. U. Böhner, J. Meisner, J. Kästner, A QUADRATICALLY-CONVERGING NUDGED ELASTIC BAND OPTIMIZER, *J. Chem. Theory Comput.* **2013**, *9*, 3498–3504 (cited on pp. 55, 61).
- [324] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, P. Gumbsch, STRUCTURAL RELAXATION MADE SIMPLE, *Phys. Rev. Lett.* **2006**, *97*, 170201 (cited on p. 56).
- [325] E. Noether, INVARIANTE VARIATIONSPROBLEME, *Nachrichten Von Ges. Wiss. Zu Gött. Math.-Phys. Kl.* **1918**, *2018*, 235–257 (cited on p. 56).
- [326] M. Melander, K. Laasonen, H. Jonsson, REMOVING EXTERNAL DEGREES OF FREEDOM FROM TRANSITION-STATE SEARCH METHODS USING QUATERNIONS, *J. Chem. Theory Comput.* **2015**, *11*, 1055–1062 (cited on p. 56).
- [327] S. K. Kearsley, ON THE ORTHOGONAL TRANSFORMATION USED FOR STRUCTURAL COMPARISONS, *Acta Cryst. A* **1989**, *45*, 208–210 (cited on p. 56).
- [328] D. L. Theobald, RAPID CALCULATION OF RMSDs USING A QUATERNION-BASED CHARACTERISTIC POLYNOMIAL, *Acta Cryst. A* **2005**, *61*, 478–480 (cited on p. 56).
- [329] S. Smidstrup, A. Pedersen, K. Stokbro, H. Jónsson, IMPROVED INITIAL GUESS FOR MINIMUM ENERGY PATH CALCULATIONS, *J. Chem. Phys.* **2014**, *140*, 214106 (cited on pp. 56, 57).

- [330] T. P. M. Goumans, C. R. A. Catlow, W. A. Brown, J. Kästner, P. Sherwood, AN EMBEDDED CLUSTER STUDY OF THE FORMATION OF WATER ON INTERSTELLAR DUST GRAINS, *Phys. Chem. Chem. Phys.* **2009**, *11*, 5431–5436 (cited on pp. 56, 61).
- [331] J. Kästner, J. M. Carr, T. W. Keal, W. Thiel, A. Wander, P. Sherwood, DL-FIND: AN OPEN-SOURCE GEOMETRY OPTIMIZER FOR ATOMISTIC SIMULATIONS, *J. Phys. Chem. A* **2009**, *113*, 11856–11865 (cited on p. 56).
- [332] T. A. Halgren, W. N. Lipscomb, THE SYNCHRONOUS-TRANSIT METHOD FOR DETERMINING REACTION PATHWAYS AND LOCATING MOLECULAR TRANSITION STATES, *Chem. Phys. Lett.* **1977**, *49*, 225–232 (cited on p. 57).
- [333] X. Zhu, K. C. Thompson, T. J. Martínez, GEODESIC INTERPOLATION FOR REACTION PATHWAYS, *J. Chem. Phys.* **2019**, *150*, 164103 (cited on pp. 57, 60).
- [334] JMOL: AN OPEN-SOURCE JAVA VIEWER FOR CHEMICAL STRUCTURES IN 3D, version 14.29, **2019**, <http://www.jmol.org/> (cited on p. 57).
- [335] P. Maragakis, S. A. Andreev, Y. Brumer, D. R. Reichman, E. Kaxiras, ADAPTIVE NUDGED ELASTIC BAND APPROACH FOR TRANSITION STATE CALCULATION, *J. Chem. Phys.* **2002**, *117*, 4651–4658 (cited on p. 58).
- [336] A. B. Birkholz, H. B. Schlegel, USING BONDING TO GUIDE TRANSITION STATE OPTIMIZATION, *J. Comput. Chem.* **2015**, *36*, 1157–1166 (cited on pp. 60, 210).
- [337] S. A. Trygubenko, D. J. Wales, A DOUBLY NUDGED ELASTIC BAND METHOD FOR FINDING TRANSITION STATES, *J. Chem. Phys.* **2004**, *120*, 2082–2094 (cited on pp. 59, 212).
- [338] L. Xie, H. Liu, W. Yang, ADAPTING THE NUDGED ELASTIC BAND METHOD FOR DETERMINING MINIMUM-ENERGY PATHS OF CHEMICAL REACTIONS IN ENZYMES, *J. Chem. Phys.* **2004**, *120*, 8039–8052 (cited on pp. 60, 212).
- [339] J. Kästner, UMBRELLA SAMPLING, *WIREs Comput. Mol. Sci.* **2011**, *1*, 932–942 (cited on p. 60).
- [340] M. U. Bohner, J. Zeman, J. Smiatek, A. Arnold, J. Kästner, NUDGED-ELASTIC BAND USED TO FIND REACTION COORDINATES BASED ON THE FREE ENERGY, *J. Chem. Phys.* **2014**, *140*, 074109 (cited on pp. 60, 61).
- [341] N. A. Zarkevich, D. D. Johnson, NUDGED-ELASTIC BAND METHOD WITH TWO CLIMBING IMAGES: FINDING TRANSITION STATES IN COMPLEX ENERGY LANDSCAPES, *J. Chem. Phys.* **2015**, *142*, 024106 (cited on p. 61).
- [342] N. R. Mathiesen, H. Jónsson, T. Vegge, J. M. García Lastra, R-NEB: ACCELERATED NUDGED ELASTIC BAND CALCULATIONS BY USE OF REFLECTION SYMMETRY, *J. Chem. Theory Comput.* **2019** (cited on p. 61).
- [343] A. A. Peterson, ACCELERATION OF SADDLE-POINT SEARCHES WITH MACHINE LEARNING, *J. Chem. Phys.* **2016**, *145*, 074106 (cited on p. 61).
- [344] O.-P. Koistinen, F. B. Dagbjartsdóttir, V. Ásgeirsson, A. Vehtari, H. Jónsson, NUDGED ELASTIC BAND CALCULATIONS ACCELERATED WITH GAUSSIAN PROCESS REGRESSION, *J. Chem. Phys.* **2017**, *147*, 152720 (cited on p. 61).
- [345] M. I. Jordan, T. M. Mitchell, MACHINE LEARNING: TRENDS, PERSPECTIVES, AND PROSPECTS, *Science* **2015**, *349*, 255–260 (cited on p. 61).
- [346] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, K. Zieba, END TO END LEARNING FOR SELF-DRIVING CARS, **2016**, arXiv: 1604.07316 (cited on p. 61).
- [347] E. Gibney, GOOGLE AI ALGORITHM MASTERS ANCIENT GAME OF GO, *Nat. News* **2016**, *529*, 445 (cited on p. 61).

- [348] A. Agrawal, A. Choudhary, PERSPECTIVE: MATERIALS INFORMATICS AND BIG DATA: REALIZATION OF THE “FOURTH PARADIGM” OF SCIENCE IN MATERIALS SCIENCE, *APL Mater.* **2016**, *4*, 053208 (cited on p. 61).
- [349] K. Mainzer, WIE BERECHENBAR IST UNSERE WELT: HERAUSFORDERUNGEN FÜR MATHEMATIK, INFORMATIK UND PHILOSOPHIE IM ZEITALTER DER DIGITALISIERUNG, essentials, *Wiesbaden*, **2018** (cited on p. 61).
- [350] M. Rupp, O. A. von Lilienfeld, K. Burke, GUEST EDITORIAL: SPECIAL TOPIC ON DATA-ENABLED THEORETICAL CHEMISTRY, *J. Chem. Phys.* **2018**, *148*, 241401 (cited on p. 61).
- [351] J. F. Rodrigues Jr, L. Florea, M. C. F. de Oliveira, D. Diamond, O. N. Oliveira Jr, A SURVEY ON BIG DATA AND MACHINE LEARNING FOR CHEMISTRY, **2019**, arXiv: 1904.10370 (cited on pp. 61, 62).
- [352] T. M. Mitchell, MACHINE LEARNING, McGraw-Hill Science, **1997** (cited on p. 61).
- [353] C. M. Bishop, PATTERN RECOGNITION AND MACHINE LEARNING, Springer, **2006** (cited on pp. 62–64, 71).
- [354] T. Hastie, R. Tibshirani, J. Friedman, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION, SECOND EDITION, 2nd ed., Springer Series in Statistics, Springer, New York, **2009** (cited on pp. 62–66, 71, 189).
- [355] G. James, D. Witten, T. Hastie, R. Tibshirani, AN INTRODUCTION TO STATISTICAL LEARNING: WITH APPLICATIONS IN R, Springer, New York, **2014** (cited on pp. 62, 63).
- [356] A. M. Turing, COMPUTING MACHINERY AND INTELLIGENCE, *Mind* **1950**, *LIX*, 433–460 (cited on p. 62).
- [357] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, MACHINE LEARNING FOR MOLECULAR AND MATERIALS SCIENCE, *Nature* **2018**, *559*, 547 (cited on p. 62).
- [358] M. Rupp, MACHINE LEARNING FOR QUANTUM MECHANICS IN A NUTSHELL, *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073 (cited on pp. 62, 67).
- [359] O. A. von Lilienfeld, QUANTUM MACHINE LEARNING IN CHEMICAL COMPOUND SPACE, *Angew. Chem. Int. Ed.* **2018**, *57*, 4164–4169 (cited on p. 62).
- [360] J. B. O. Mitchell, MACHINE LEARNING METHODS IN CHEMOINFORMATICS: MACHINE LEARNING METHODS IN CHEMOINFORMATICS, *WIREs Comput. Mol. Sci.* **2014**, *4*, 468–481 (cited on p. 62).
- [361] T. Mueller, A. G. Kusne, R. Ramprasad, MACHINE LEARNING IN MATERIALS SCIENCE: RECENT PROGRESS AND EMERGING APPLICATIONS, in *Reviews in Computational Chemistry, Vol. 29*, (Eds.: A. L. Parrill, K. B. Lipkowitz), first, John Wiley & Sons, Inc., **2016**, pp. 186–273 (cited on p. 62).
- [362] R. B. Jadrich, B. A. Lindquist, T. M. Truskett, RECENT ADVANCES IN ACCELERATED DISCOVERY THROUGH MACHINE LEARNING AND STATISTICAL INFERENCE, **2017**, arXiv: 1706.05405 (cited on p. 62).
- [363] A. C. Mater, M. L. Coote, DEEP LEARNING IN CHEMISTRY, *J. Chem. Inf. Model.* **2019** (cited on p. 62).
- [364] N. J. Browning, R. Ramakrishnan, O. A. von Lilienfeld, U. Roethlisberger, GENETIC OPTIMIZATION OF TRAINING SETS FOR IMPROVED MACHINE LEARNING MODELS OF MOLECULAR PROPERTIES, *J. Phys. Chem. Lett.* **2017**, 1351–1359 (cited on p. 62).
- [365] M. S. Jørgensen, M. N. Groves, B. Hammer, COMBINING EVOLUTIONARY ALGORITHMS WITH CLUSTERING TOWARD RATIONAL GLOBAL STRUCTURE OPTIMIZATION AT THE ATOMIC SCALE, *J. Chem. Theory Comput.* **2017**, *13*, 1486–1493 (cited on pp. 62, 110).

- [366] M. S. Jørgensen, U. F. Larsen, K. W. Jacobsen, B. Hammer, EXPLORATION VERSUS EXPLOITATION IN GLOBAL ATOMISTIC STRUCTURE OPTIMIZATION, *J. Phys. Chem. A* **2018** (cited on p. 62).
- [367] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, SCIKIT-LEARN: MACHINE LEARNING IN PYTHON, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830 (cited on pp. 63, 64, 72).
- [368] E. Schubert, A. Zimek, ELKI: A LARGE OPEN-SOURCE LIBRARY FOR DATA ANALYSIS - ELKI RELEASE 0.7.5 "HEIDELBERG", **2019**, arXiv: 1902.03616 (cited on p. 63).
- [369] F. Chollet *et al.*, KERAS, **2015**, <https://keras.io> (visited on 06/19/2019) (cited on p. 63).
- [370] A. K. Jain, DATA CLUSTERING: 50 YEARS BEYOND K-MEANS, *Pattern Recognit. Lett.* **2010**, *31*, 651–666 (cited on pp. 63, 65).
- [371] A. K. Jain, M. N. Murty, P. J. Flynn, DATA CLUSTERING: A REVIEW, *ACM Comput. Surv.* **1999**, *31*, 264–323 (cited on pp. 63, 65).
- [372] J. Behler, M. Parrinello, GENERALIZED NEURAL-NETWORK REPRESENTATION OF HIGH-DIMENSIONAL POTENTIAL-ENERGY SURFACES, *Phys. Rev. Lett.* **2007**, *98*, 146401 (cited on p. 63).
- [373] J. Behler, NEURAL NETWORK POTENTIAL-ENERGY SURFACES IN CHEMISTRY: A TOOL FOR LARGE-SCALE SIMULATIONS, *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930 (cited on p. 63).
- [374] J. Behler, REPRESENTING POTENTIAL ENERGY SURFACES BY HIGH-DIMENSIONAL NEURAL NETWORK POTENTIALS, *J. Phys. Condens. Matter* **2014**, *26*, 183001 (cited on p. 63).
- [375] J. Behler, CONSTRUCTING HIGH-DIMENSIONAL NEURAL NETWORK POTENTIALS: A TUTORIAL REVIEW, *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050 (cited on pp. 63, 67).
- [376] J. Behler, FIRST PRINCIPLES NEURAL NETWORK POTENTIALS FOR REACTIVE SIMULATIONS OF LARGE MOLECULAR AND CONDENSED SYSTEMS, *Angew. Chem. Int. Ed.* **2017**, *56*, 12828–12840 (cited on p. 63).
- [377] J. Behler, ATOM-CENTERED SYMMETRY FUNCTIONS FOR CONSTRUCTING HIGH-DIMENSIONAL NEURAL NETWORK POTENTIALS, *J. Chem. Phys.* **2011**, *134*, 074106 (cited on pp. 63, 67).
- [378] M. Ceriotti, UNSUPERVISED MACHINE LEARNING IN ATOMISTIC SIMULATIONS, BETWEEN PREDICTIONS AND UNDERSTANDING, *J. Chem. Phys.* **2019**, *150*, 150901 (cited on pp. 64, 67).
- [379] L. Van Der Maaten, E. Postma, J. Van den Herik, DIMENSIONALITY REDUCTION: A COMPARATIVE REVIEW, *J. Mach. Learn. Res.* **2009**, *10*, 66–71 (cited on p. 64).
- [380] J. D. Leeuw, I. J. R. Barra, F. Brodeau, G. Romier, B. V. Cutsem (eds), APPLICATIONS OF CONVEX ANALYSIS TO MULTIDIMENSIONAL SCALING, in *Recent Developments in Statistics*, North Holland Publishing Company, **1977**, pp. 133–146 (cited on p. 64).
- [381] J. de Leeuw, P. Mair, MULTIDIMENSIONAL SCALING USING MAJORIZATION: SMACOF IN R, *J. Stat. Soft.* **2009**, *31* (cited on p. 64).
- [382] A. P. Bartók, R. Kondor, G. Csányi, ON REPRESENTING CHEMICAL ENVIRONMENTS, *Phys. Rev. B* **2013**, *87*, 184115 (cited on p. 67).
- [383] M. J. Willatt, F. Musil, M. Ceriotti, ATOM-DENSITY REPRESENTATIONS FOR MACHINE LEARNING, *J. Chem. Phys.* **2019**, *150*, 154110 (cited on p. 67).
- [384] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, FAST AND ACCURATE MODELING OF MOLECULAR ATOMIZATION ENERGIES WITH MACHINE LEARNING, *Phys. Rev. Lett.* **2012**, *108*, 058301 (cited on p. 67).

- [385] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, MACHINE LEARNING PREDICTIONS OF MOLECULAR PROPERTIES: ACCURATE MANY-BODY POTENTIALS AND NONLOCALITY IN CHEMICAL SPACE, *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331 (cited on p. 67).
- [386] B. Huang, O. A. von Lilienfeld, COMMUNICATION: UNDERSTANDING MOLECULAR REPRESENTATIONS IN MACHINE LEARNING: THE ROLE OF UNIQUENESS AND TARGET SIMILARITY, *J. Chem. Phys.* **2016**, *145*, 161102 (cited on p. 67).
- [387] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, FOURIER SERIES OF ATOMIC RADIAL DISTRIBUTION FUNCTIONS: A MOLECULAR FINGERPRINT FOR MACHINE LEARNING MODELS OF QUANTUM CHEMICAL PROPERTIES, *Int. J. Quantum Chem.* **2015**, *115*, 1084–1093 (cited on p. 67).
- [388] H. Huo, M. Rupp, UNIFIED REPRESENTATION FOR MACHINE LEARNING OF MOLECULES AND CRYSTALS, **2017**, arXiv: 1704.06439 (cited on p. 67).
- [389] J. M. Dieterich, GENETIC ALGORITHMS IN THEORETICAL CHEMISTRY – DEVELOPMENT AND APPLICATIONS OF A GENERAL PURPOSE FRAMEWORK, *PhD Thesis*, Christian-Albrechts-University, Kiel, **2010** (cited on pp. 69, 73, 109).
- [390] J. M. Dieterich, B. Hartke, OGOLEM: GLOBAL CLUSTER STRUCTURE OPTIMISATION FOR ARBITRARY MIXTURES OF FLEXIBLE MOLECULES. A MULTISCALING, OBJECT-ORIENTED APPROACH, *Mol. Phys.* **2010**, *108*, 279–291 (cited on pp. 69, 95, 109).
- [391] J. M. Dieterich, B. Hartke, COMPOSITION-INDUCED STRUCTURAL TRANSITIONS IN MIXED LENNARD-JONES CLUSTERS: GLOBAL REPARAMETRIZATION AND OPTIMIZATION, *J. Comput. Chem.* **2011**, *32*, 1377–1385 (cited on p. 69).
- [392] R. P. Gupta, LATTICE RELAXATION AT A METAL SURFACE, *Phys. Rev. B* **1981**, *23*, 6265–6270 (cited on p. 70).
- [393] J. M. Dieterich, U. Gerstel, J.-M. Schröder, B. Hartke, AGGREGATION OF KANAMYCIN A: DIMER FORMATION WITH PHYSIOLOGICAL CATIONS, *J. Mol. Model.* **2011**, *17*, 3195–3207 (cited on p. 70).
- [394] J. M. Dieterich, B. Hartke, EMPIRICAL REVIEW OF STANDARD BENCHMARK FUNCTIONS USING EVOLUTIONARY GLOBAL OPTIMIZATION, *Appl. Math.* **2012**, *3*, 1552–1564 (cited on p. 70).
- [395] R. M. Forck, J. M. Dieterich, C. C. Pradzynski, A. L. Huchting, R. A. Mata, T. Zeuch, STRUCTURAL DIVERSITY IN SODIUM DOPED WATER TRIMERS, *Phys. Chem. Chem. Phys.* **2012**, *14*, 9054–9057 (cited on p. 70).
- [396] M. Sierka, SYNERGY BETWEEN THEORY AND EXPERIMENT IN STRUCTURE RESOLUTION OF LOW-DIMENSIONAL OXIDES, *Prog. Surf. Sci.* **2010**, *85*, 398–434 (cited on p. 70).
- [397] S. Heiles, R. L. Johnston, GLOBAL OPTIMIZATION OF CLUSTERS USING ELECTRONIC STRUCTURE METHODS, *Int. J. Quantum Chem.* **2013**, *113*, 2091–2109 (cited on pp. 70, 117).
- [398] J. M. Dieterich, G. H. Clever, R. A. Mata, A PUSH-AND-PULL MODEL FOR ALLOSTERIC ANION BINDING IN CAGE COMPLEXES, *Phys. Chem. Chem. Phys.* **2012**, *14*, 12746–12749 (cited on p. 70).
- [399] M. Frank, J. M. Dieterich, S. Freye, R. A. Mata, G. H. Clever, RELATIVE ANION BINDING AFFINITY IN A SERIES OF INTERPENETRATED COORDINATION CAGES, *Dalton Trans.* **2013**, *42*, 15906–15910 (cited on p. 70).
- [400] Y. Li, B. Hartke, ASSESSING SOLVATION EFFECTS ON CHEMICAL REACTIONS WITH GLOBALLY OPTIMIZED SOLVENT CLUSTERS, *ChemPhysChem* **2013**, *14*, 2678–2686 (cited on pp. 70, 229).

- [401] U. Buck, C. C. Pradzynski, T. Zeuch, J. M. Dieterich, B. Hartke, A SIZE RESOLVED INVESTIGATION OF LARGE WATER CLUSTERS, *Phys. Chem. Chem. Phys.* **2014**, *16*, 6859 (cited on p. 70).
- [402] J. M. Dieterich, B. Hartke, A GRAPH-BASED SHORT-CUT TO LOW-ENERGY STRUCTURES, *J. Comput. Chem.* **2014**, *35*, 1618–1620 (cited on p. 70).
- [403] J. M. Dieterich, B. Hartke, OBSERVABLE-TARGETING GLOBAL CLUSTER STRUCTURE OPTIMIZATION, *Phys. Chem. Chem. Phys.* **2015**, *17*, 11958–11961 (cited on p. 70).
- [404] J. M. Dieterich, B. Hartke, ERROR-SAFE, PORTABLE, AND EFFICIENT EVOLUTIONARY ALGORITHMS IMPLEMENTATION WITH HIGH SCALABILITY, *J. Chem. Theory Comput.* **2016**, *12*, 5226–5233 (cited on pp. 70, 72, 109, 116).
- [405] F. Spenke, K. Balzer, S. Frick, B. Hartke, J. M. Dieterich, MALLEABLE PARALLELISM WITH MINIMAL EFFORT FOR MAXIMAL THROUGHPUT AND MAXIMAL HARDWARE LOAD, *Comput. Theor. Chem.* **2019**, *1151*, 72–77 (cited on pp. 70, 109, 116).
- [406] J. M. Dieterich, B. Hartke, IMPROVED CLUSTER STRUCTURE OPTIMIZATION: HYBRIDIZING EVOLUTIONARY ALGORITHMS WITH LOCAL HEAT PULSES, *Inorganics* **2017**, *5*, 64 (cited on pp. 71, 79, 156).
- [407] B. G. del Rio, J. M. Dieterich, E. A. Carter, GLOBALLY-OPTIMIZED LOCAL PSEUDOPOTENTIALS FOR (ORBITAL-FREE) DENSITY FUNCTIONAL THEORY SIMULATIONS OF LIQUIDS AND SOLIDS, *J. Chem. Theory Comput.* **2017**, *13*, 3684–3695 (cited on pp. 71, 72).
- [408] C. Witt, J. M. Dieterich, B. Hartke, CLUSTER STRUCTURES INFLUENCED BY INTERACTION WITH A SURFACE, *Phys. Chem. Chem. Phys.* **2018**, *20*, 15661–15670 (cited on p. 71).
- [409] A. Freibert, J. M. Dieterich, B. Hartke, EXPLORING SELF-ORGANIZATION OF MOLECULAR TETHER MOLECULES ON A GOLD SURFACE BY GLOBAL STRUCTURE OPTIMIZATION, *J. Comput. Chem.* **2019** (cited on p. 71).
- [410] H. M. Aktulga, S. A. Pandit, A. C. T. van Duin, A. Y. Grama, REACTIVE MOLECULAR DYNAMICS: NUMERICAL METHODS AND ALGORITHMIC TECHNIQUES, *SIAM J. Sci. Comp.* **2012**, *34*, C1–C23 (cited on pp. 71, 72, 95, 115).
- [411] J. C. Fogarty, H. M. Aktulga, A. Y. Grama, A. C. T. van Duin, S. A. Pandit, A REACTIVE MOLECULAR DYNAMICS SIMULATION OF THE SILICA-WATER INTERFACE, *J. Chem. Phys.* **2010**, *132*, 174704 (cited on pp. 71, 72, 95, 115).
- [412] A. Danial, CLOC – COUNT LINES OF CODE, version 1.74, <https://github.com/AlDanial/cloc> (visited on 03/11/2019) (cited on p. 71).
- [413] J. M. Dieterich, OGOLEM.ORG Homepage, <https://www.ogolem.org/> (visited on 04/15/2018) (cited on pp. 72, 111).
- [414] G. S. Fanourgakis, S. S. Xantheas, DEVELOPMENT OF TRANSFERABLE INTERACTION POTENTIALS FOR WATER. V. EXTENSION OF THE FLEXIBLE, POLARIZABLE, THOLE-TYPE MODEL POTENTIAL (TTM3-F, v. 3.0) TO DESCRIBE THE VIBRATIONAL SPECTRA OF WATER CLUSTERS AND LIQUID WATER, *J. Chem. Phys.* **2008**, *128*, 074506 (cited on p. 72).
- [415] T. Oliphant, A GUIDE TO NUMPY, USA: Trelgol Publishing, **2006**, <http://www.numpy.org/> (cited on p. 72).
- [416] T. E. Oliphant, PYTHON FOR SCIENTIFIC COMPUTING, *Comput. Sci. Eng.* **2007**, *9*, 10–20 (cited on p. 72).
- [417] J. D. Hunter, MATPLOTLIB: A 2D GRAPHICS ENVIRONMENT, *Comput. Sci. Eng.* **2007**, *9*, 90–95 (cited on p. 72).
- [418] W. McKinney, DATA STRUCTURES FOR STATISTICAL COMPUTING IN PYTHON, in Proceedings of the 9th Python in Science Conference, **2010**, pp. 51–56 (cited on p. 72).

- [419] M. Waskom, SEABORN: STATISTICAL DATA VISUALIZATION, version 0.9.0, **2018**, <https://seaborn.pydata.org/> (cited on p. 72).
- [420] S. Seabold, J. Perktold, STATSMODELS: ECONOMETRIC AND STATISTICAL MODELING WITH PYTHON, in Proceedings of the 9th Python in Science Conference, **2010**, pp. 57–61 (cited on p. 72).
- [421] T. Williams, C. Kelley, many others, GNU PLOT 5.2: AN INTERACTIVE PLOTTING PROGRAM, **2018** (cited on p. 72).
- [422] B. Bandow, B. Hartke, LARGER WATER CLUSTERS WITH EDGES AND CORNERS ON THEIR WAY TO ICE: STRUCTURAL TRENDS ELUCIDATED WITH AN IMPROVED PARALLEL EVOLUTIONARY ALGORITHM, *J. Phys. Chem. A* **2006**, *110*, 5809–5822 (cited on pp. 72, 73, 109).
- [423] Y. Ge, J. D. Head, FAST GLOBAL OPTIMIZATION OF SIXHY CLUSTERS: NEW MUTATION OPERATORS IN THE CLUSTER GENETIC ALGORITHM, *Chem. Phys. Lett.* **2004**, *398*, 107–112 (cited on p. 73).
- [424] D. C. Lonie, E. Zurek, XTALOPT: AN OPEN-SOURCE EVOLUTIONARY ALGORITHM FOR CRYSTAL STRUCTURE PREDICTION, *Comput. Phys. Commun.* **2011**, *182*, 372–387 (cited on p. 73).
- [425] A. Shayeghi, D. Götz, J. B. A. Davis, R. Schäfer, R. L. Johnston, POOL-BCGA: A PARALLELISED GENERATION-FREE GENETIC ALGORITHM FOR THE AB INITIO GLOBAL OPTIMISATION OF NANOALLOY CLUSTERS, *Phys. Chem. Chem. Phys.* **2015**, *17*, 2104–2112 (cited on p. 73).
- [426] L. B. Vilhelmsen, B. Hammer, A GENETIC ALGORITHM FOR FIRST PRINCIPLES GLOBAL STRUCTURE OPTIMIZATION OF SUPPORTED NANO STRUCTURES, *J. Chem. Phys.* **2014**, *141*, 044711 (cited on p. 73).
- [427] J. A. Vargas, F. Buendía, M. R. Beltrán, NEW AuN (N = 27–30) LOWEST ENERGY CLUSTERS OBTAINED BY MEANS OF AN IMPROVED DFT–GENETIC ALGORITHM METHODOLOGY, *J. Phys. Chem. C* **2017**, *121*, 10982–10991 (cited on p. 73).
- [428] M. Jäger, R. Schäfer, R. L. Johnston, GIGA: A VERSATILE GENETIC ALGORITHM FOR FREE AND SUPPORTED CLUSTERS AND NANOPARTICLES IN THE PRESENCE OF LIGANDS, *Nanoscale* **2019**, *11*, 9042–9052 (cited on p. 73).
- [429] W. Thiel, MNDO2005, version 7.0, Max-Planck-Institute für Kohlenforschung an der Ruhr, Germany, **2005** (cited on p. 76).
- [430] F. Neese, THE ORCA PROGRAM SYSTEM, *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78 (cited on p. 76).
- [431] D. H. Wolpert, W. G. Macready, NO FREE LUNCH THEOREMS FOR OPTIMIZATION, *Trans. Evol. Comput.* **1997**, *1*, 67–82 (cited on pp. 76, 147).
- [432] A. Klamt, G. Schüürmann, COSMO: A NEW APPROACH TO DIELECTRIC SCREENING IN SOLVENTS WITH EXPLICIT EXPRESSIONS FOR THE SCREENING ENERGY AND ITS GRADIENT, *J. Chem. Soc. Perkin Trans. 2* **1993**, 799–805 (cited on p. 86).
- [433] M. Dittner, B. Hartke, GLOBALLY OPTIMAL CATALYTIC FIELDS – INVERSE DESIGN OF ABSTRACT EMBEDDINGS FOR MAXIMUM REACTION RATE ACCELERATION, *J. Chem. Theory Comput.* **2018**, *14*, 3547–3564 (cited on pp. 86, 88, 128, 287).
- [434] A. Langer, Java Generics FAQs, <http://www.angelikalanger.com/GenericsFAQ/JavaGenericsFAQ.html> (visited on 05/13/2019) (cited on p. 109).
- [435] Oracle, Java Language and Virtual Machine Specifications, <https://docs.oracle.com/javase/specs/> (visited on 05/14/2019) (cited on p. 109).

- [436] G. M. Amdahl, VALIDITY OF THE SINGLE PROCESSOR APPROACH TO ACHIEVING LARGE SCALE COMPUTING CAPABILITIES, in Proceedings of the April 18-20, 1967, Spring Joint Computer Conference on - AFIPS '67 (Spring), The April 18-20, 1967, Spring Joint Computer Conference, ACM Press, Atlantic City, New Jersey, **1967**, p. 483 (cited on p. 109).
- [437] J. L. Gustafson, REEVALUATING AMDAHL'S LAW, *Commun. ACM* **1988**, *31*, 532–533 (cited on p. 109).
- [438] M. D. Hill, M. R. Marty, AMDAHL'S LAW IN THE MULTICORE ERA, *Computer* **2008**, *41*, 33–38 (cited on p. 110).
- [439] Dong Hyuk Woo, H.-H. Lee, EXTENDING AMDAHL'S LAW FOR ENERGY-EFFICIENT COMPUTING IN THE MANY-CORE ERA, *Computer* **2008**, *41*, 24–31 (cited on p. 110).
- [440] B. Göetz, T. Peierls, J. Bloch, J. Bowbeer, D. Holmes, D. Lea, JAVA CONCURRENCY IN PRACTICE, 1st edition, Addison-Wesley Professional, **2006** (cited on pp. 110, 114).
- [441] J. Evans, *A Scalable Concurrent Malloc(3) Implementation for FreeBSD*, **2006**, <http://jemalloc.net/> (visited on 05/12/2019) (cited on p. 110).
- [442] S. Liang, THE JAVA™ NATIVE INTERFACE, PROGRAMMER'S GUIDE AND SPECIFICATION, Addison-Wesley, **1999** (cited on p. 111).
- [443] J. Bloch, EFFECTIVE JAVA (THE JAVA SERIES), 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, USA, **2008** (cited on p. 111).
- [444] B. Eckel, THINKING IN JAVA: THE DEFINITIVE INTRODUCTION TO OBJECT-ORIENTED PROGRAMMING IN THE LANGUAGE OF THE WORLD WIDE WEB, 4th edition, Prentice Hall, Upper Saddle River, NJ, **2006** (cited on p. 111).
- [445] E. Gamma, R. Helm, R. E. Johnson, J. Vlissides, DESIGN PATTERNS. ELEMENTS OF REUSABLE OBJECT-ORIENTED SOFTWARE. 1st edition, Prentice Hall, Reading, Mass, **1994** (cited on p. 111).
- [446] E. Freeman, E. Freeman, B. Bates, K. Sierra, HEAD FIRST DESIGN PATTERNS, O' Reilly & Associates, Inc., **2004** (cited on p. 111).
- [447] S. J. Metsker, W. C. Wake, DESIGN PATTERNS IN JAVA, Addison Wesley, **2006** (cited on p. 111).
- [448] R. L. Johnston, EVOLVING BETTER NANOPARTICLES: GENETIC ALGORITHMS FOR OPTIMISING CLUSTER GEOMETRIES, *Dalton Trans.* **2003**, 4193–4207 (cited on p. 117).
- [449] M. Dittner, B. Hartke, CONQUERING THE HARD CASES OF LENNARD-JONES CLUSTERS WITH SIMPLE RECIPES, *Comput. Theor. Chem.* **2017**, *1107*, 7–13 (cited on pp. 118, 287).
- [450] J. C. Culberson, ON THE FUTILITY OF BLIND SEARCH: AN ALGORITHMIC VIEW OF “NO FREE LUNCH”, *Evol. Comput.* **1998**, *6*, 109–127 (cited on p. 147).
- [451] Y.-C. Ho, D. L. Pepyne, SIMPLE EXPLANATION OF THE NO-FREE-LUNCH THEOREM AND ITS IMPLICATIONS, *J. Optim. Theory Appl.* **2002**, *115*, 549–570 (cited on p. 147).
- [452] F. Weigend, R. Ahlrichs, BALANCED BASIS SETS OF SPLIT VALENCE, TRIPLE ZETA VALENCE AND QUADRUPLE ZETA VALENCE QUALITY FOR H TO RN: DESIGN AND ASSESSMENT OF ACCURACY, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305 (cited on p. 158).
- [453] B. H. Besler, K. M. Merz, P. A. Kollman, ATOMIC CHARGES DERIVED FROM SEMIEMPIRICAL METHODS, *J. Comput. Chem.* **1990**, *11*, 431–439 (cited on p. 166).
- [454] I. T. Suydam, C. D. Snow, V. S. Pande, S. G. Boxer, ELECTRIC FIELDS AT THE ACTIVE SITE OF AN ENZYME: DIRECT COMPARISON OF EXPERIMENT WITH THEORY, *Science* **2006**, *313*, 200–204 (cited on p. 171).

- [455] C. T. Liu, J. P. Layfield, R. J. Stewart, J. B. French, P. Hanoian, J. B. Asbury, S. Hammes-Schiffer, S. J. Benkovic, PROBING THE ELECTROSTATICS OF ACTIVE SITE MICROENVIRONMENTS ALONG THE CATALYTIC CYCLE FOR ESCHERICHIA COLI DIHYDROFOLATE REDUCTASE, *J. Am. Chem. Soc.* **2014**, *136*, 10349–10360 (cited on pp. 171, 173, 179, 183).
- [456] M. Reppert, A. Tokmakoff, COMPUTATIONAL AMIDE I 2D IR SPECTROSCOPY AS A PROBE OF PROTEIN STRUCTURE AND DYNAMICS, *Annu. Rev. Phys. Chem.* **2016**, *67*, 359–386 (cited on p. 171).
- [457] S. D. Fried, S. G. Boxer, MEASURING ELECTRIC FIELDS AND NONCOVALENT INTERACTIONS USING THE VIBRATIONAL STARK EFFECT, *Acc. Chem. Res.* **2015**, *48*, 998–1006 (cited on pp. 171, 178).
- [458] J. Villà, A. Warshel, ENERGETICS AND DYNAMICS OF ENZYMATIC REACTIONS, *J. Phys. Chem. B* **2001**, *105*, 7887–7907 (cited on p. 172).
- [459] S. Shaik, D. Mandal, R. Ramanan, ORIENTED ELECTRIC FIELDS AS FUTURE SMART REAGENTS IN CHEMISTRY, *Nat. Chem.* **2016**, *8*, 1091–1098 (cited on pp. 177, 181, 206–208, 212, 226, 227).
- [460] S. Ciampi, N. Darwish, H. M. Aitken, I. Díez-Pérez, M. L. Coote, HARNESSING ELECTROSTATIC CATALYSIS IN SINGLE MOLECULE, ELECTROCHEMICAL AND CHEMICAL SYSTEMS: A RAPIDLY GROWING EXPERIMENTAL TOOL BOX, *Chem. Soc. Rev.* **2018**, *47*, 5146–5164 (cited on pp. 177, 179).
- [461] S. Shaik, S. P. de Visser, D. Kumar, EXTERNAL ELECTRIC FIELD WILL CONTROL THE SELECTIVITY OF ENZYMATIC-LIKE BOND ACTIVATIONS, *J. Am. Chem. Soc.* **2004**, *126*, 11746–11749 (cited on p. 178).
- [462] K.-B. Cho, H. Hirao, H. Chen, M. A. Carvajal, S. Cohen, E. Derat, W. Thiel, S. Shaik, COMPOUND I IN HEME THIOLATE ENZYMES: A COMPARATIVE QM/MM STUDY, *J. Phys. Chem. A* **2008**, *112*, 13128–13138 (cited on pp. 178, 208).
- [463] R. Meir, H. Chen, W. Lai, S. Shaik, ORIENTED ELECTRIC FIELDS ACCELERATE DIELS–ALDER REACTIONS AND CONTROL THE ENDO/EXO SELECTIVITY, *ChemPhysChem* **2010**, *11*, 301–310 (cited on pp. 178, 180, 181, 206–208, 212).
- [464] K. Bhattacharyya, S. Karmakar, A. Datta, EXTERNAL ELECTRIC FIELD CONTROL: DRIVING THE REACTIVITY OF METAL-FREE AZIDE–ALKYNE CLICK REACTIONS, *Phys. Chem. Chem. Phys.* **2017**, *19*, 22482–22486 (cited on p. 178).
- [465] Z. Wang, D. Danovich, R. Ramanan, S. Shaik, ORIENTED-EXTERNAL ELECTRIC FIELDS CREATE ABSOLUTE ENANTIOSELECTIVITY IN DIELS–ALDER REACTIONS: IMPORTANCE OF THE MOLECULAR DIPOLE MOMENT, *J. Am. Chem. Soc.* **2018**, *140*, 13350–13359 (cited on pp. 178, 180, 212, 227).
- [466] R. Ramanan, D. Danovich, D. Mandal, S. Shaik, CATALYSIS OF METHYL TRANSFER REACTIONS BY ORIENTED EXTERNAL ELECTRIC FIELDS: ARE GOLD–THIOLATE LINKERS INNOCENT?, *J. Am. Chem. Soc.* **2018**, *140*, 4354–4362 (cited on p. 178).
- [467] S. Shaik, P. C. Hiberty, THE CHEMIST’S GUIDE TO VALENCE BOND THEORY, first edition, Wiley-Interscience, Hoboken, NJ, **2008** (cited on p. 178).
- [468] S. Shaik, A. Shurki, VALENCE BOND DIAGRAMS AND CHEMICAL REACTIVITY, *Angew. Chem. Int. Ed.* **1999**, *38*, 586–625 (cited on p. 178).
- [469] G. Cassone, F. Pietrucci, F. Saija, F. Guyot, A. M. Saitta, ONE-STEP ELECTRIC-FIELD DRIVEN METHANE AND FORMALDEHYDE SYNTHESIS FROM LIQUID METHANOL, *Chem. Sci.* **2017**, *8*, 2329–2336 (cited on p. 178).
- [470] G. Cassone, J. Spöner, J. E. Spöner, F. Pietrucci, A. M. Saitta, F. Saija, SYNTHESIS OF (D)-ERYTHROSE FROM GLYCOLALDEHYDE AQUEOUS SOLUTIONS UNDER ELECTRIC FIELD, *Chem. Commun.* **2018**, *54*, 3211–3214 (cited on p. 178).

- [471] G. Cassone, F. Pietrucci, F. Saija, F. Guyot, J. Sponer, J. E. Sponer, A. M. Saitta, NOVEL ELECTROCHEMICAL ROUTE TO CLEANER FUEL DIMETHYL ETHER, *Sci. Rep.* **2017**, *7*, 6901 (cited on p. 178).
- [472] A. C. Aragonès, N. L. Haworth, N. Darwish, S. Ciampi, N. J. Bloomfield, G. G. Wallace, I. Diez-Perez, M. L. Coote, ELECTROSTATIC CATALYSIS OF A DIELS–ALDER REACTION, *Nature* **2016**, *531*, 88–91 (cited on pp. 179, 180, 212).
- [473] L. Zhang, E. Laborda, N. Darwish, B. B. Noble, J. H. Tyrell, S. Pluczyk, A. P. Le Brun, G. G. Wallace, J. Gonzalez, M. L. Coote, S. Ciampi, ELECTROCHEMICAL AND ELECTROSTATIC CLEAVAGE OF ALKOXYAMINES, *J. Am. Chem. Soc.* **2018**, *140*, 766–774 (cited on p. 179).
- [474] M. Klinska, L. M. Smith, G. Gryn'ova, M. G. Banwell, M. L. Coote, EXPERIMENTAL DEMONSTRATION OF PH-DEPENDENT ELECTROSTATIC CATALYSIS OF RADICAL REACTIONS, *Chem. Sci.* **2015**, *6*, 5623–5627 (cited on p. 179).
- [475] H. M. Aitken, M. L. Coote, CAN ELECTROSTATIC CATALYSIS OF DIELS–ALDER REACTIONS BE HARNESSSED WITH PH-SWITCHABLE CHARGED FUNCTIONAL GROUPS?, *Phys. Chem. Chem. Phys.* **2018**, *20*, 10671–10676 (cited on pp. 179, 180, 212).
- [476] F. Che, J. T. Gray, S. Ha, N. Kruse, S. L. Scott, J.-S. McEwen, ELUCIDATING THE ROLES OF ELECTRIC FIELDS IN CATALYSIS: A PERSPECTIVE, *ACS Catal.* **2018**, *8*, 5153–5174 (cited on p. 180).
- [477] P. C. Hiberty, C. Megret, L. Song, W. Wu, S. Shaik, BARRIERS OF HYDROGEN ABSTRACTION VS HALOGEN EXCHANGE: AN EXPERIMENTAL MANIFESTATION OF CHARGE-SHIFT BONDING, *J. Am. Chem. Soc.* **2006**, *128*, 2836–2843 (cited on p. 181).
- [478] J. A. Pople, G. A. Segal, APPROXIMATE SELF-CONSISTENT MOLECULAR ORBITAL THEORY. II. CALCULATIONS WITH COMPLETE NEGLECT OF DIFFERENTIAL OVERLAP, *J. Chem. Phys.* **1965**, *43*, S136–S151 (cited on p. 206).
- [479] J. Tomasi, B. Mennucci, R. Cammi, QUANTUM MECHANICAL CONTINUUM SOLVATION MODELS, *Chem. Rev.* **2005**, *105*, 2999–3094 (cited on p. 207).
- [480] D. Behrens, GLOBAL OPTIMIZATION OF ABSTRACT CATALYSTS: ON THE ROAD TO APPLICATION, Master's Thesis, Christian-Albrechts-University, Kiel, **2019** (cited on pp. 213–219).
- [481] O. Khersonsky, D. Röthlisberger, O. Dym, S. Albeck, C. J. Jackson, D. Baker, D. S. Tawfik, EVOLUTIONARY OPTIMIZATION OF COMPUTATIONALLY DESIGNED ENZYMES: KEMP ELIMINASES OF THE KE07 SERIES, *J. Mol. Biol.* **2010**, *396*, 1025–1042 (cited on p. 214).
- [482] O. Khersonsky, D. Röthlisberger, A. M. Wollacott, P. Murphy, O. Dym, S. Albeck, G. Kiss, K. N. Houk, D. Baker, D. S. Tawfik, OPTIMIZATION OF THE IN-SILICO-DESIGNED KEMP ELIMINASE KE70 BY COMPUTATIONAL DESIGN AND DIRECTED EVOLUTION, *J. Mol. Biol.* **2011**, *407*, 391–412 (cited on p. 214).
- [483] O. Khersonsky, G. Kiss, D. Röthlisberger, O. Dym, S. Albeck, K. N. Houk, D. Baker, D. S. Tawfik, BRIDGING THE GAPS IN DESIGN METHODOLOGIES BY EVOLUTIONARY OPTIMIZATION OF THE STABILITY AND PROFICIENCY OF DESIGNED KEMP ELIMINASE KE59, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 10358–10363 (cited on p. 214).
- [484] T. Kinnunen, K. Laasonen, THE OXIDATIVE ADDITION AND MIGRATORY 1,1-INSERTION IN THE MONSANTO AND CATIVA PROCESSES. A DENSITY FUNCTIONAL STUDY OF THE CATALYTIC CARBONYLATION OF METHANOL, *J. Mol. Struct. THEOCHEM* **2001**, *542*, 273–288 (cited on p. 217).
- [485] L. Cavallo, M. Solà, A THEORETICAL STUDY OF STERIC AND ELECTRONIC EFFECTS IN THE RHODIUM-CATALYZED CARBONYLATION REACTIONS, *J. Am. Chem. Soc.* **2001**, *123*, 12294–12302 (cited on p. 217).

- [486] T. Kinnunen, K. Laasonen, DFT-STUDIES OF *CIS*- AND *TRANS*-[Rh(CO)₂X₂]⁺ (X=PH₃, PF₃, PCl₃, PBr₃, PI₃ OR P(CH₃)₃) AND OXIDATIVE ADDITION OF CH₃I TO THEM, *J. Organomet. Chem.* **2003**, *665*, 150–155 (cited on p. 217).
- [487] M. Feliz, Z. Freixa, P. W. N. M. van Leeuwen, C. Bo, REVISITING THE METHYL IODIDE OXIDATIVE ADDITION TO RHODIUM COMPLEXES: A DFT STUDY OF THE ACTIVATION PARAMETERS, *Organometallics* **2005**, *24*, 5718–5723 (cited on p. 217).
- [488] R. H. Crabtree, THE ORGANOMETALLIC CHEMISTRY OF THE TRANSITION METALS, John Wiley & Sons, Inc., Hoboken, **2005** (cited on p. 219).
- [489] D. M. Behrens, CURRENT WORK, Christian-Albrechts University, Kiel, **2019** (cited on p. 226).
- [490] M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, HEURISTICS-GUIDED EXPLORATION OF REACTION MECHANISMS, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722 (cited on p. 231).
- [491] T. J. Martínez, AB INITIO REACTIVE COMPUTER AIDED MOLECULAR DESIGN, *Acc. Chem. Res.* **2017**, *50*, 652–656 (cited on p. 231).
- [492] L.-P. Wang, R. T. McGibbon, V. S. Pande, T. J. Martinez, AUTOMATED DISCOVERY AND REFINEMENT OF REACTIVE MOLECULAR DYNAMICS PATHWAYS, *J. Chem. Theory Comput.* **2016**, *12*, 638–649 (cited on p. 231).
- [493] D. Rappoport, C. J. Galvin, D. Y. Zubarev, A. Aspuru-Guzik, COMPLEX CHEMICAL REACTION NETWORKS FROM HEURISTICS-AIDED QUANTUM CHEMISTRY, *J. Chem. Theory Comput.* **2014**, *10*, 897–907 (cited on p. 231).
- [494] G. N. Simm, A. C. Vaucher, M. Reiher, EXPLORATION OF REACTION PATHWAYS AND CHEMICAL TRANSFORMATION NETWORKS, *J. Phys. Chem. A* **2019**, *123*, 385–399 (cited on p. 231).
- [495] F. De Vleeschouwer, P. Geerlings, F. D. Proft, MOLECULAR PROPERTY OPTIMIZATIONS WITH BOUNDARY CONDITIONS THROUGH THE BEST FIRST SEARCH SCHEME, *ChemPhysChem* **2016**, *17*, 141–1424 (cited on p. 232).
- [496] T. D. Huan, A. Mannodi-Kanakkithodi, R. Ramprasad, ACCELERATED MATERIALS PROPERTY PREDICTIONS AND DESIGN USING MOTIF-BASED FINGERPRINTS, *Phys. Rev. B* **2015**, *92*, 014106 (cited on p. 232).
- [497] M. Springborg, S. Kohaut, Y. Dong, K. Huwig, MIXED SI-GE CLUSTERS, SOLAR-ENERGY HARVESTING, AND INVERSE-DESIGN METHODS, *Comput. Theor. Chem.* **2017**, *1107*, 14–22 (cited on p. 232).
- [498] J. L. Teunissen, F. De Proft, F. De Vleeschouwer, ACCELERATION OF INVERSE MOLECULAR DESIGN BY USING PREDICTIVE TECHNIQUES, *J. Chem. Inf. Model.* **2019** (cited on p. 232).
- [499] F. Spenke, CURRENT WORK, Christian-Albrechts University, Kiel, **2019** (cited on p. 232).

Globally Optimal Catalyst Optimization

A.1 Operator Benchmarks for the Genetic Algorithm

Complementary to the studies of Section 6.3.1 on p. 147, benchmark runs for the same input settings as compiled in Table 6.1 are given for the GOCAT sizes $N_{\text{Ch}} = 5$ and $N_{\text{Ch}} = 20$ in the following. For the latter case, the flexibility of the GOCATs and, as a result, also the search space was increased by enforcing only a minimal distances of $r_{ij} \geq r_{\text{min}} = 0.1 \text{ \AA}$ between two charges, i and j —all other GOCAT models for $N_{\text{Ch}} \in \{5, 10\}$ have $r_{ij} \geq 1.0 \text{ \AA}$ instead, as usual.

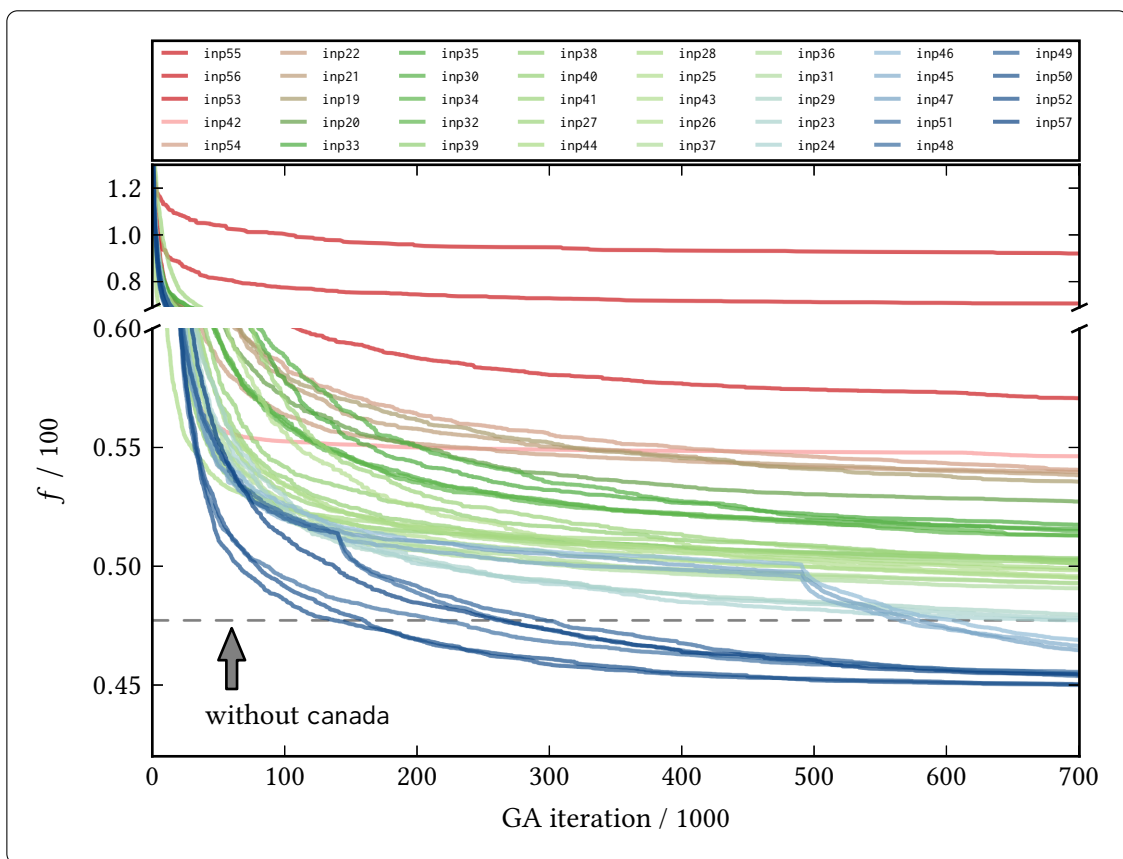


Fig. A.1: $N_{\text{Ch}} = 5$: Averaged fitness progression. This plot is the same as Fig. 6.1 on p. 152 but shows a benchmark for the smaller GOCAT size $N_{\text{Ch}} = 5$. See Table 6.1 on p. 151 for the input setting definitions. The fitness values given in that same Table do not belong to these runs with $N_{\text{Ch}} = 5$.

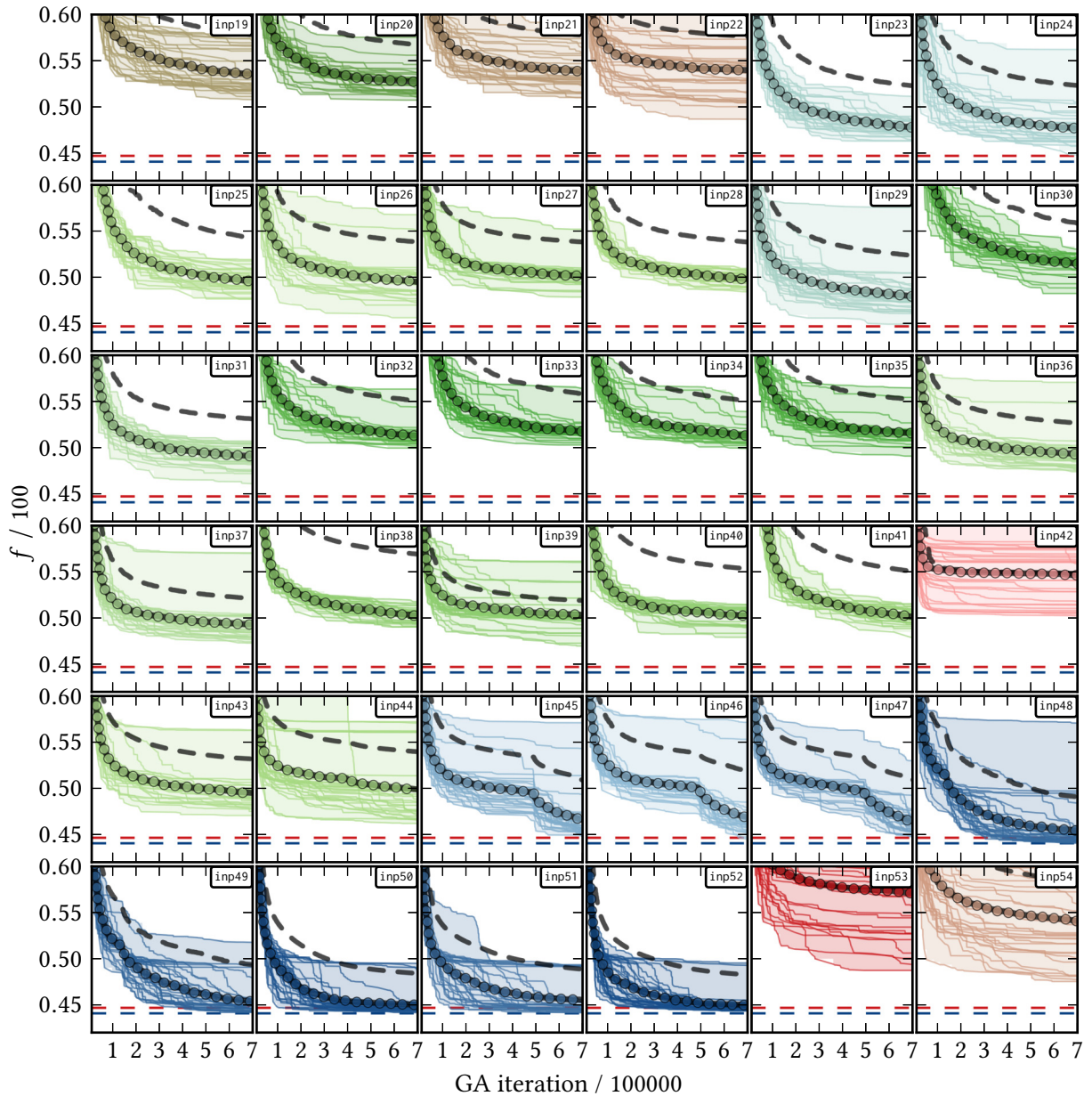


Fig. A.2: $N_{Ch} = 5$: Multiplot of 36 settings of Fig. A.1 on the previous page. This plot is the same as Fig. 6.2 on p. 153 but shows a benchmark for the smaller **GOCAT** size $N_{Ch} = 5$. See Table 6.1 on p. 151 for the input setting definitions. The fitness values given in that same Table do not belong to these runs with $N_{Ch} = 5$.

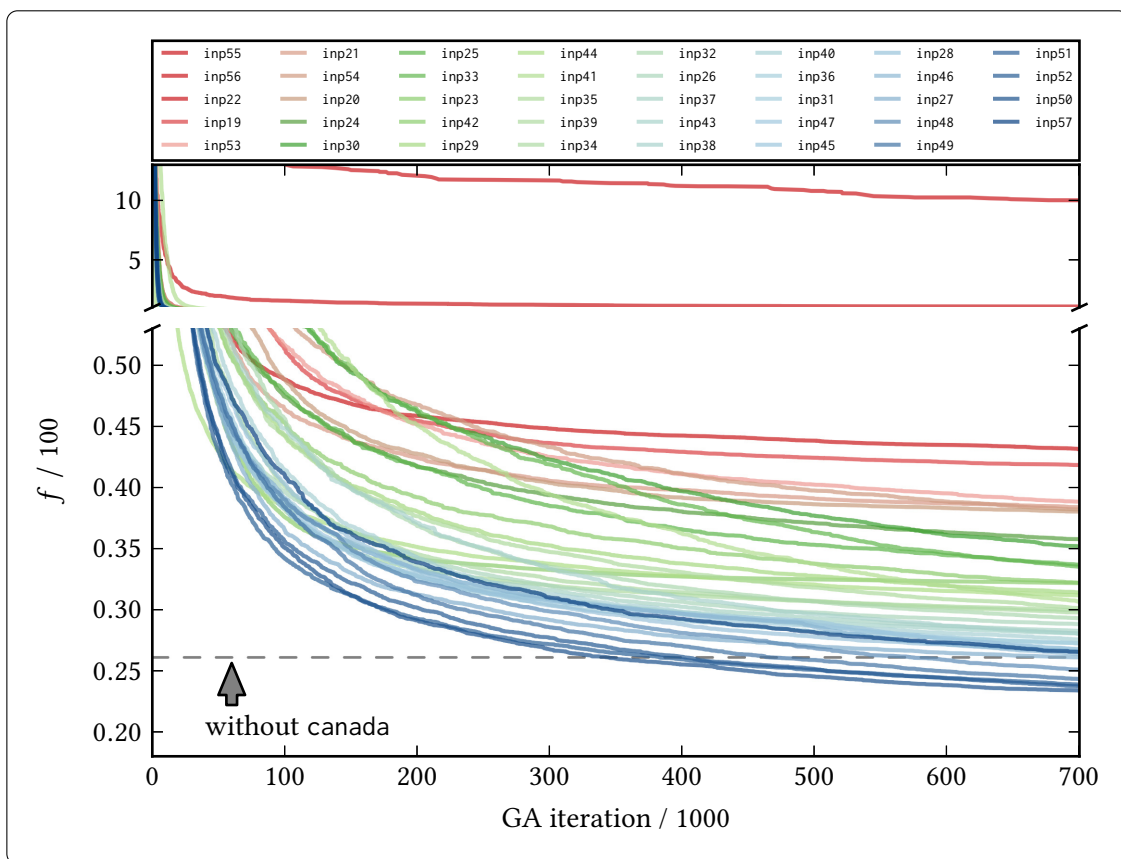


Fig. A.3: $N_{Ch} = 20$: Averaged fitness progression. This plot is the same as Fig. 6.1 on p. 152 but shows a benchmark for the larger GOCAT size $N_{Ch} = 20$. See Table 6.1 on p. 151 for the input setting definitions. The fitness values given in that same Table do not belong to these runs with $N_{Ch} = 20$.

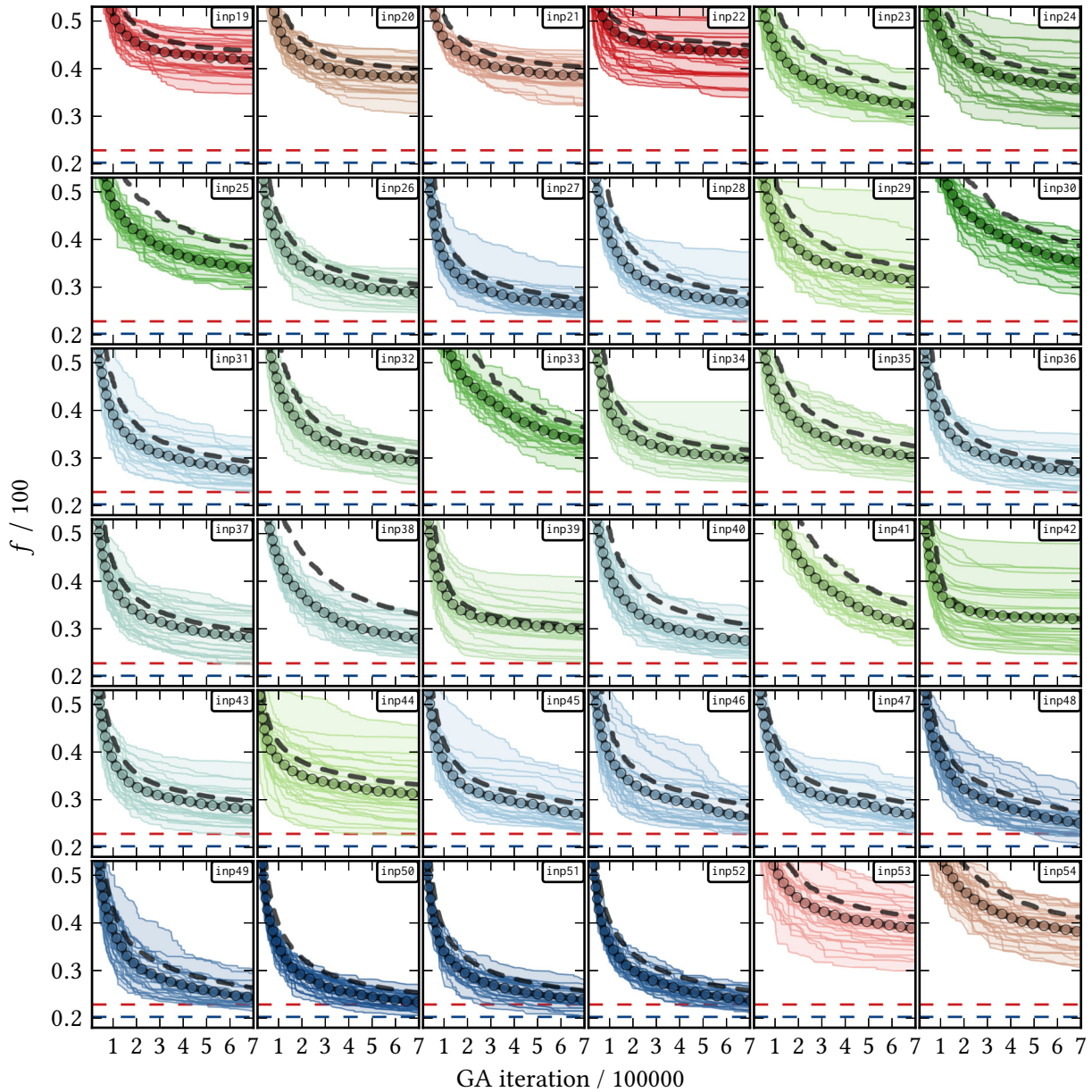


Fig. A.4: $N_{Ch} = 20$: Multiplot of 36 settings of Fig. A.3. This plot is the same as Fig. 6.2 on p. 153 but shows a benchmark for the larger GOCAT size $N_{Ch} = 20$. See Table 6.1 on p. 151 for the input setting definitions. The fitness values given in that same Table do not belong to these runs with $N_{Ch} = 20$.

A.2 Electric Fields for the Menshutkin Reaction

For the sake of completeness, ESP/EF plots for the implicit solvent model COSMO around the Menshutkin reaction path as well as some optimized GOCATs are given in Figs. A.5 to A.7 on pp. 270–272. These plots are complementary to the COSMO reference discussions in the publication of Section 6.2, particularly of Section 3.4 “ $N_{\text{Ch}} = 10$ Stabilizing COSMO Path” on pp. 138ff. The COSMO comparisons can indicate the differences between the frame-wise changing COSMO stabilization of each frame vs. the overall static GOCAT for all frames at once. The field along the Cl–C–N axis can be seen in all cases, which was discussed in Section 6.3.3.

For the COSMO case in Fig. A.5, NH_3 is rotated off the direct line to the CH_3 in the R frame. The reaction field of the solvent for the TS and P structures are then fully uniform along the Cl–C–N axis in Figs. A.6(a) and A.6(b). However, the R frame is stabilized by a positive potential at the free electron pair of N and a negative one on the other side. Evidently, the implicit solvent leads to an inhomogeneous field. In contrast, in the GOCAT optimization, such asymmetric frames are not included using the pristine gas phase and just *one* overall embedding for all frames (*cf.* Fig. A.7(a)); this GOCAT embedding is optimized with regard to optimal barrier decrease instead of a stabilization energy at each frame, as it is the case for COSMO. In Fig. A.7(b), catalyzing the COSMO MEP instead of the gas phase path, the final EF agrees better with the pristine (no GOCAT) COSMO cases (Fig. A.6(a)), as expected.

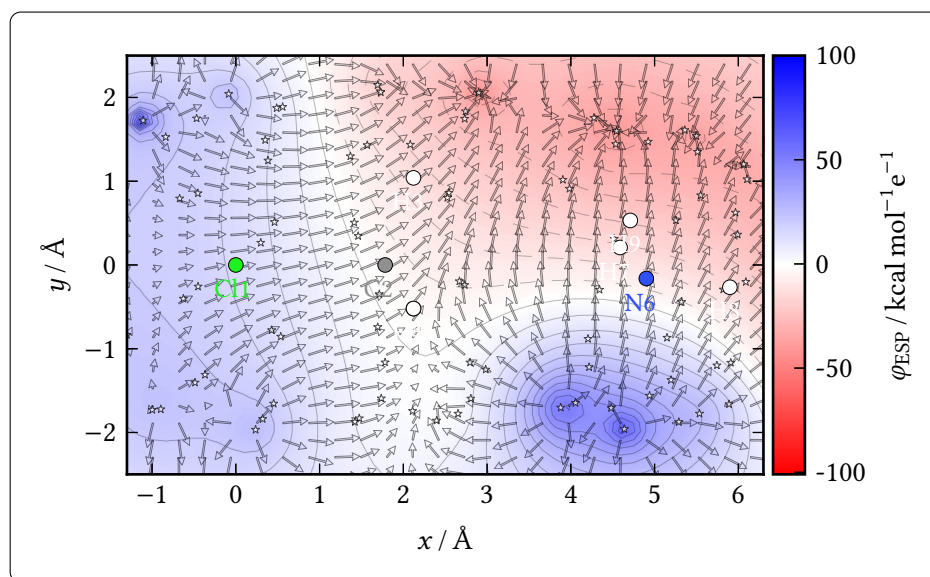
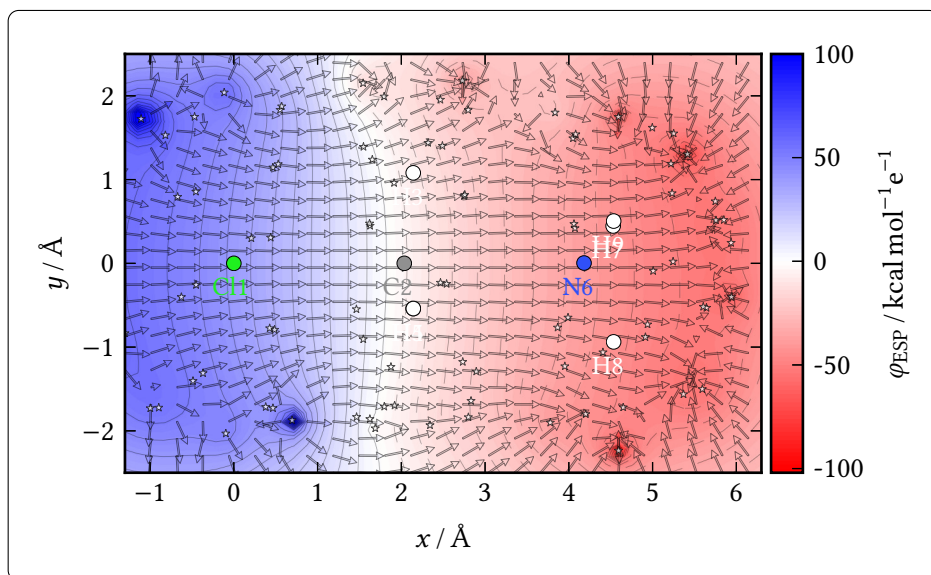
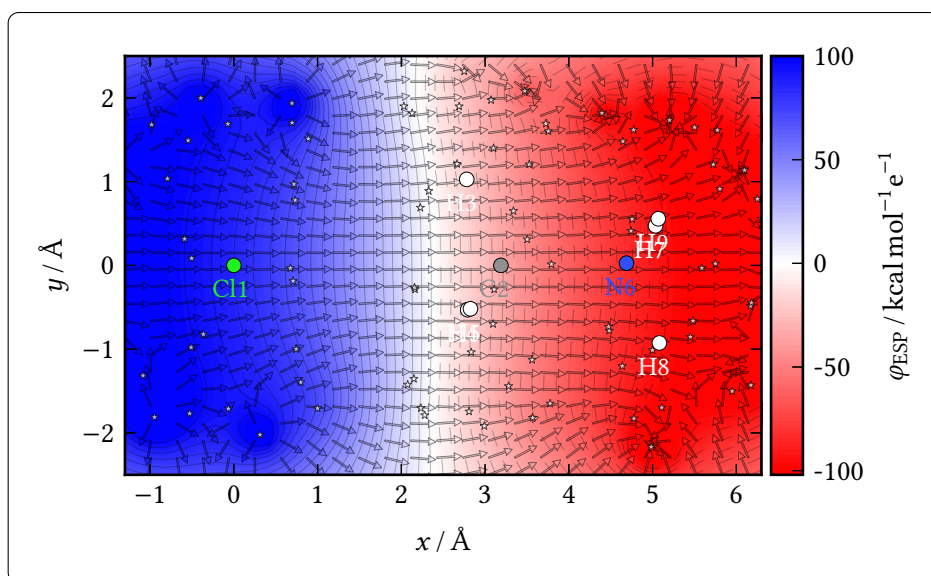


Fig. A.5: ESP and EF within the implicit solvent model COSMO (no GOCAT) around the Menshutkin reaction for the R frame as discussed in Section 6.2 on pp. 138ff. ESP, φ_{ESP} , and its EF as the negative gradient plotted as arrows are shown, where the EF is projected onto the plane of Cl1, C2 and H3 for the corresponding frame. Gradients are clipped when becoming too large for visualization purposes because of the singularities at the Coulomb charge centers. Contour lines are drawn for each $\Delta\varphi_{\text{ESP}} = 5 \text{ kcal mol}^{-1} \text{ e}^{-1}$.

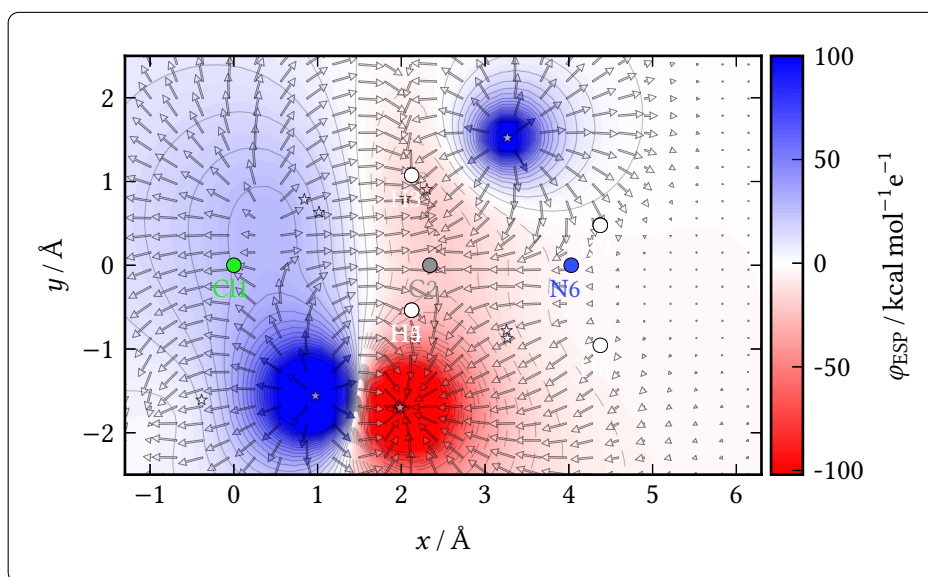


(a) TS frame in COSMO

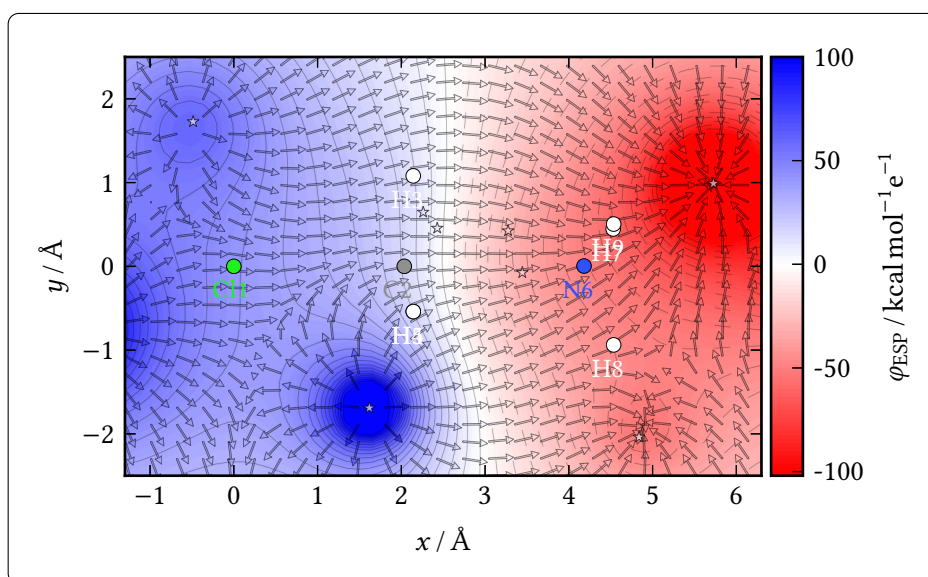


(b) P frame in COSMO

Fig. A.6: Both the TS and P frames within the implicit solvent model COSMO (no GOCAT) are given, complementary to Fig. A.5. For plotting details, see Fig. A.5.



(a) TS frame of the $r\theta$ GOCAT for $N_{\text{Ch}} = 10$



(b) TS frame of the $r\theta$ GOCAT for $N_{\text{Ch}} = 10$ using the COSMO MEP

Fig. A.7: In Fig. (a), the highly symmetric best GOCAT $r\theta$ is shown for $N_{\text{Ch}} = 10$ on a vdW surface (compare with $r\theta$ in Figs. 6.7(a) and 6.7(b) on p. 167). In Fig. (b), the best GOCAT $r\theta$ is shown for $N_{\text{Ch}} = 10$ on a vdW surface, but starting with the COSMO structures, cf. “Figure 11” on p. 140. For plotting details, see Fig. A.5.

A.3 Clustering for the Diels–Alder Reaction and Further Data

Some discussions in the Thesis were based on separate chunks of the data, *i.e.*, clusters, after applying **HC** to the complete separate **GOCAT** databases of a particular model. This part is complementary to the Section 7.4.1. Dendrograms for **HC** are given in Figs. A.8 and A.9 on the current page and on the following page and 2D **MDS** projections in Figs. A.10 and A.11 on p. 275 and on p. 276, all of these for the static models with $N_{\text{Ch}} \in \{10, 81\}$. Furthermore for $N_{\text{Ch}} = 10$ (**vdW**, static), all individuals in the best clusters are illustrated as a superposition of all charges. For the fitness function including the **R** stabilization restraint ($\Delta E_{\text{R}} \leq 0 \text{ kcal mol}^{-1}$), this is shown in Fig. A.12 on p. 276; for the fitness function variant without such an additional restraint, all charges are shown in Fig. A.14 on p. 277. Note that without the restraint, $\Delta E_{\text{R}} > 0 \text{ kcal mol}^{-1}$ clearly is developed in the final **GOCATs** for the **DA** reaction, seen in Fig. A.13 on p. 277, although the former cases with **R** stabilization are still existing in the search space. However, they have not yet emerged because the **R** destabilization leads to a better fitness and dominates in this fitness function. There is no further fitness incentive in the case without the restraint, in neither direction for the energy of **R**.

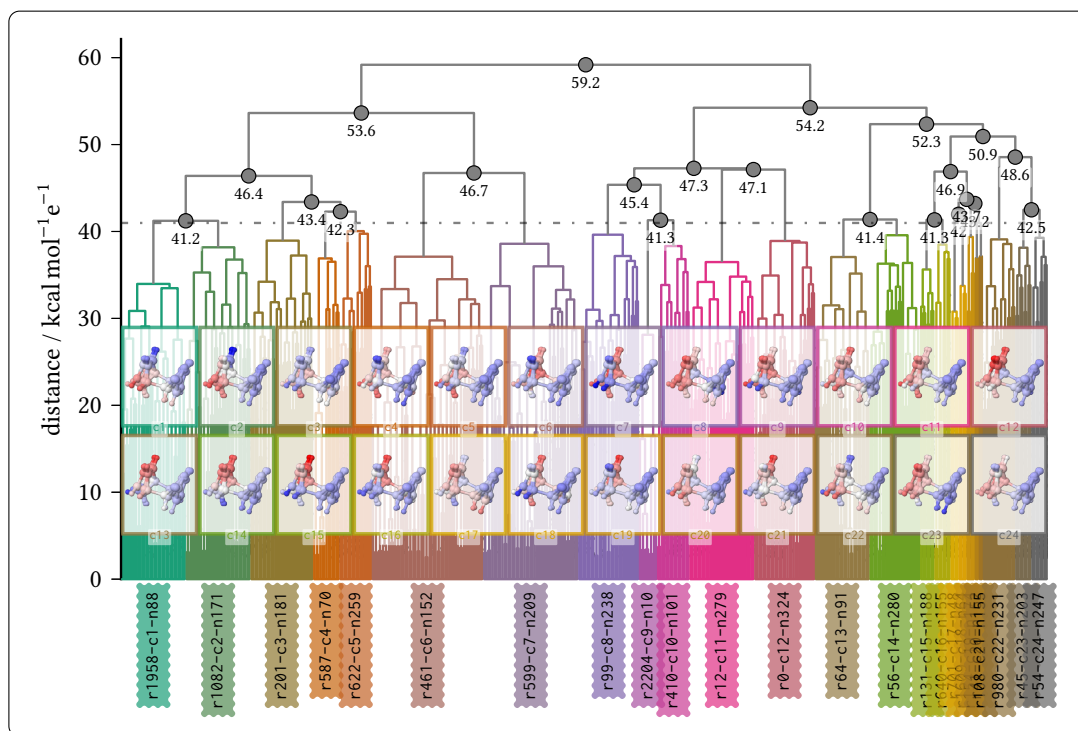


Fig. A.8: $N_{\text{Ch}} = 81$ (sphere, static): Dendrogram of agglomerative **HC** for the *endo* **DA** reaction used in Figs. 7.3 and 7.6 on p. 186 and on p. 191, chunked into 24 clusters, cutting at a distance of $d = 41.0 \text{ kcal mol}^{-1} e^{-1}$. Also, the three superposed frames (**R**, **TS**, **P**) are shown for a few selected clusters, whose φ_{ESP} values are colored from red to blue for the interval $[-20.7, +20.7] \text{ kcal mol}^{-1} e^{-1}$ based on the *averaged* **ESP** within each cluster.

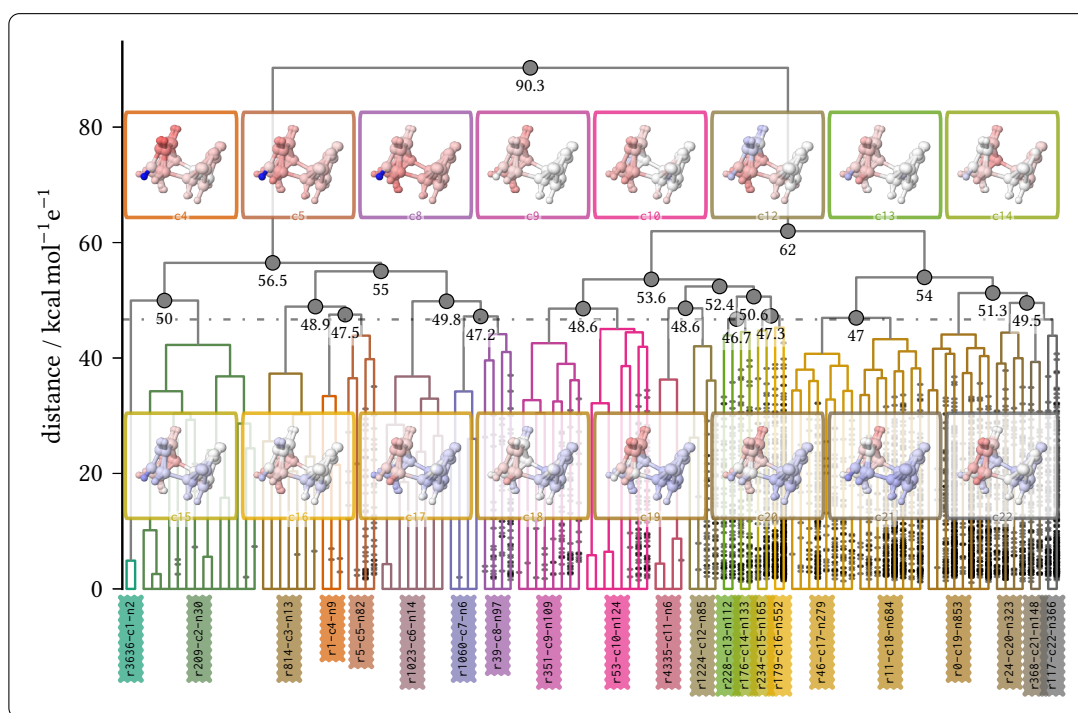


Fig. A.9: $N_{\text{Ch}} = 10$ (vdW, static): Dendrogram of agglomerative HC for the *endo* DA reaction used in Fig. 7.7 on p. 192, chunked into 22 clusters, cutting at a distance of $d = 46.7 \text{ kcal mol}^{-1} \text{ e}^{-1}$. Also, the three superposed frames (R, TS, P) are shown for a few selected clustered, whose φ_{ESP} values are colored from red to blue for the interval $[-48.4, 48.4] \text{ kcal mol}^{-1} \text{ e}^{-1}$ based on the *averaged* ESP within each cluster.

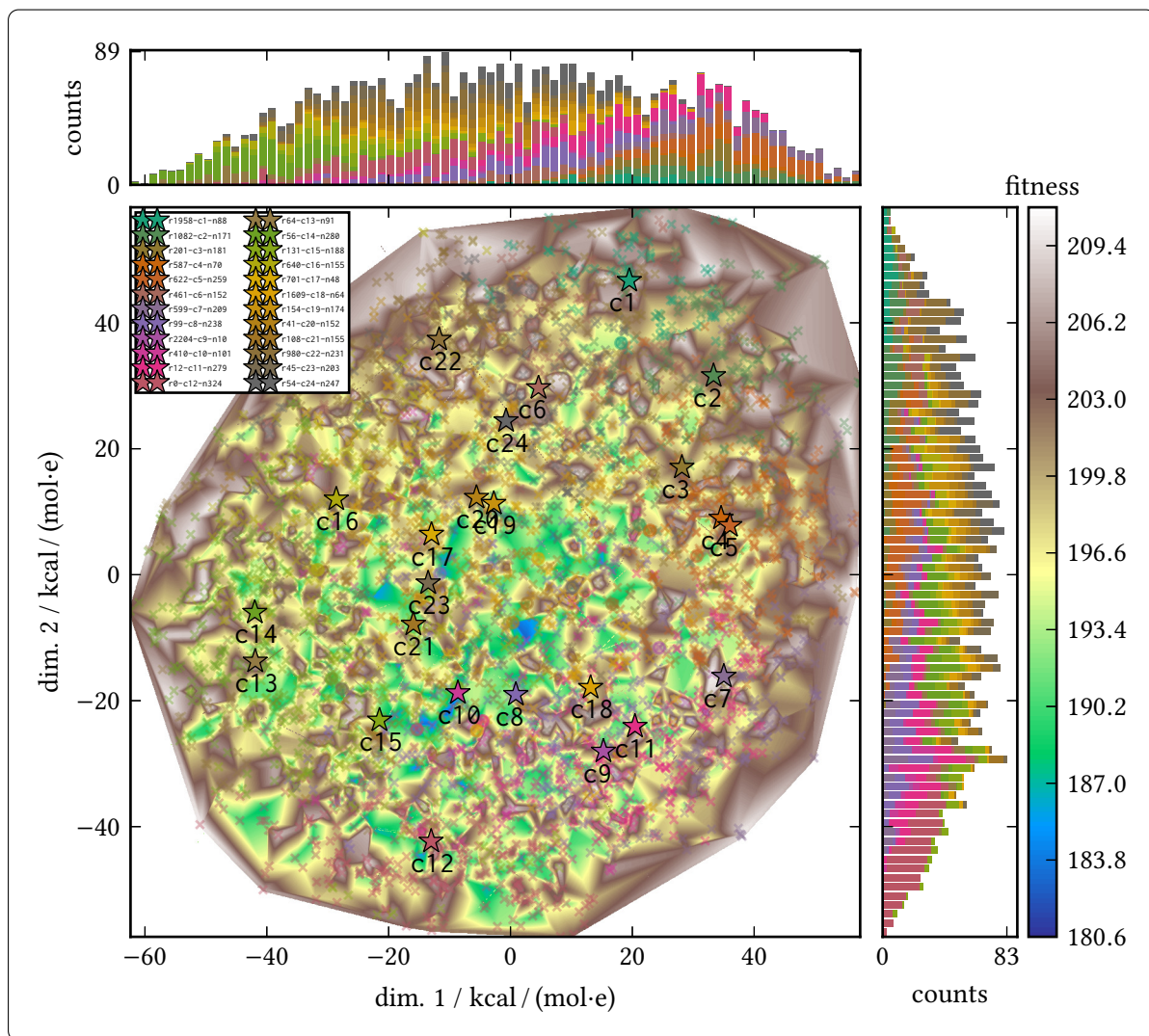


Fig. A.10: $N_{\text{Ch}} = 81$ (sphere, static): 2D MDS projection of the data cut into the 24 clusters of Fig. A.8. Stars show the mean position of each cluster plotted over a linearly interpolated fitness surface of the database.

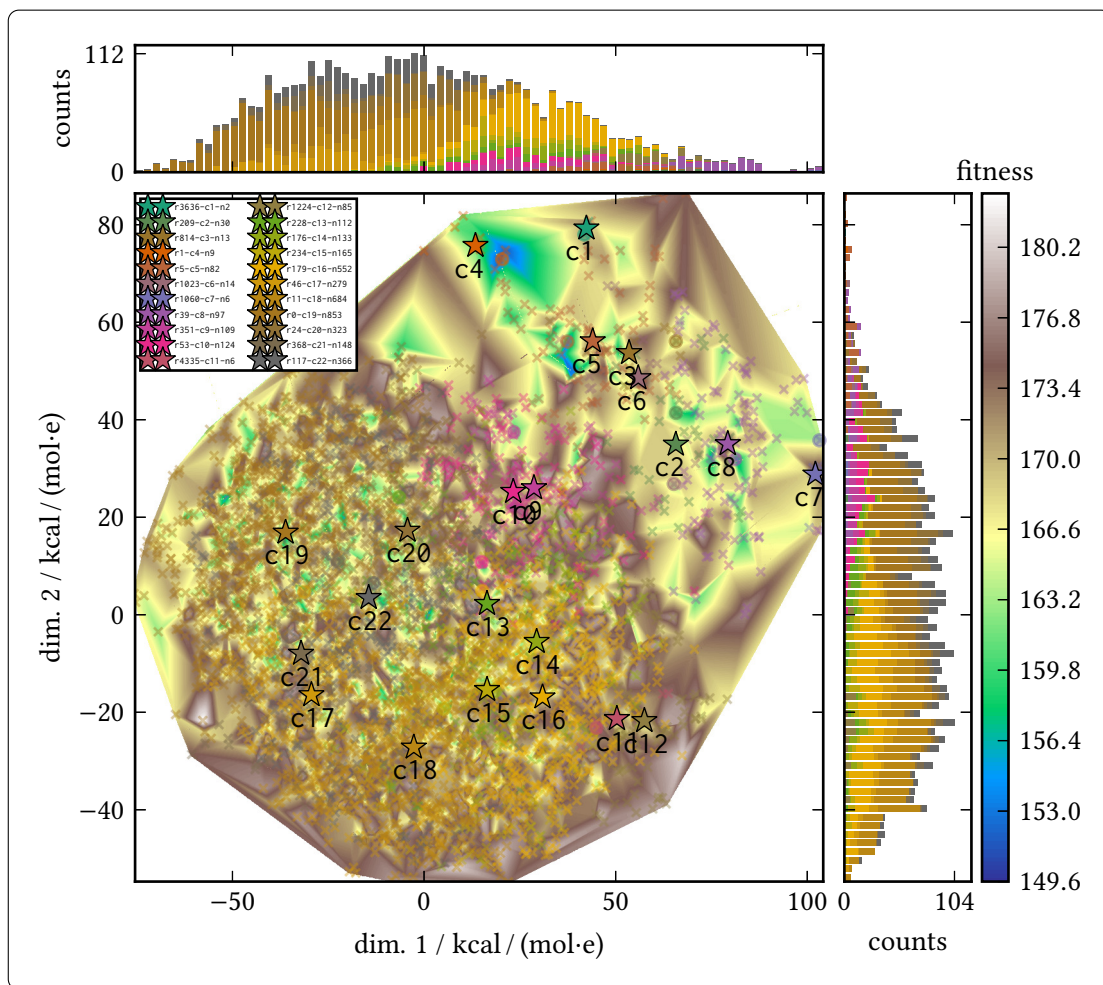


Fig. A.11: $N_{\text{Ch}} = 10$ (vdW, static): 2D MDS projection of the data cut into the 22 clusters of Fig. A.9. Stars show the mean position of each cluster plotted over a linearly interpolated fitness surface of the database.

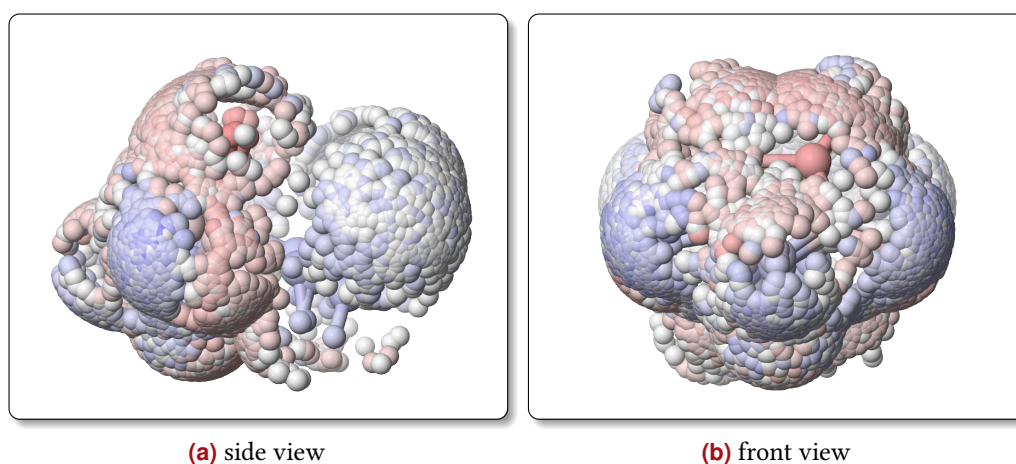


Fig. A.12: $N_{\text{Ch}} = 10$ (vdW, static): All 10 · 853 charges superposed for c19, complementary to Figs. A.9 and A.11 and to Figs. 7.7 and 7.8 on p. 192 and on p. 193.

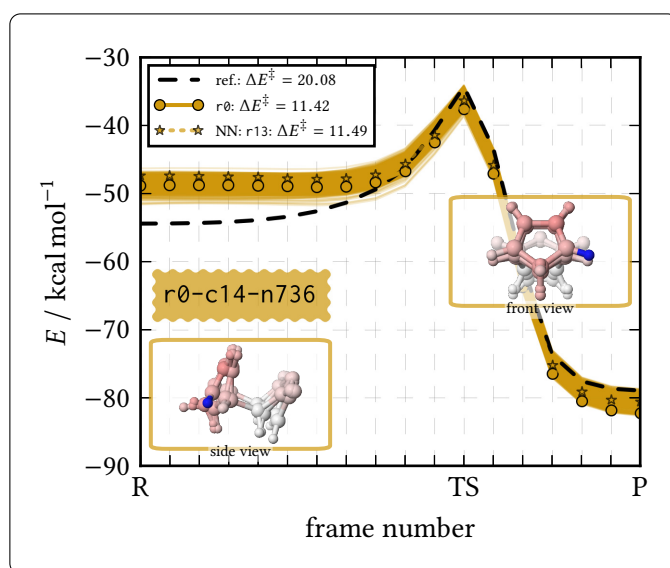


Fig. A.13: $N_{\text{Ch}} = 10$ (vdW, static, $\Delta E_{\text{R}} > 0$ allowed): Reaction energy profiles of GOCATs for the *endo* DA reaction for the best cluster, c14, with the best GOCAT without the $\Delta E_{\text{R}} \leq 0$ kcal mol⁻¹ restraint leading to R destabilization as main barrier decrease mechanism. Notice the very positively embedded H-atom from the R frame that serves as handle to destabilize this structure. Whether this is seen as a meaningful mechanism is quite another topic. Assuming linearity of the superposition of possible influences of the ESP for the sake of this argument, such local impacts are often possible and can also be used as handle *after* TSS in the other GOCATs (cf. Section 7.4) for shifting up the energy of the R side—probably a case of overfitting. φ_{ESP} values are colored from red to blue for the interval $[-48.7, 48.7]$ kcal mol⁻¹ e⁻¹ based on the averaged ESP within the cluster. Compare with Fig. A.14.

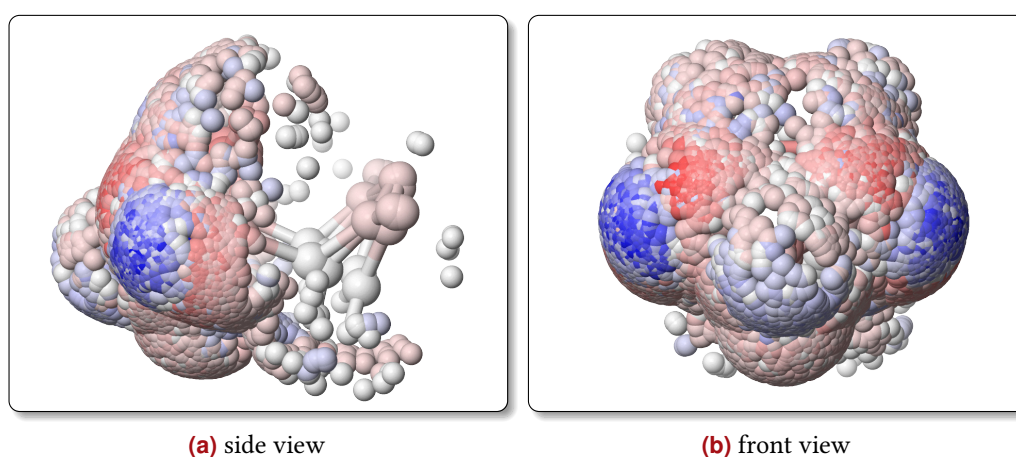


Fig. A.14: $N_{\text{Ch}} = 10$ (vdW, static, $\Delta E_{\text{R}} > 0$ allowed): All $10 \cdot 736$ charges superposed for c14, $q_i \in [-1.0, 1.0]$ e. Compare with Fig. A.13.

A.4 Correlations for the Diels–Alder Reaction

A few correlations of GOCAT properties complementary to Section 7.4 are given here. Already in Fig. 7.3 on p. 186, the energies of another cluster c3 were shown. This cluster does not contain the best rank but it shows the highest correlation of the energy barrier, ΔE^\ddagger , with respect to the property $\Delta\varphi_{\text{ESP-TS-}z\text{-pl}}$. However, ΔE_{TS} is *not* correlated with the same feature. At the same time, $\Delta E_{\text{R}} < 0 \text{ kcal mol}^{-1}$ holds because of the restraint in the fitness function. As a result, there must be candidate solutions that *do* stabilize the TS frame and also *do* stabilize the R frame but the R frame to a smaller extent in order to generate a barrier decrease. This exactly is the correlation between ΔE_{R} and ΔE_{TS} that is indeed present. Yet, there is a variation of the $\Delta\varphi_{\text{ESP-TS-}z\text{-pl}}$ in both the positive domain—that is the expected field along the “reaction axis”—and the *negative* domain. Although this cluster is one with a seemingly strong correlation along this *z*-plane, the GOCAT effects are thus *not* simply explainable by TSS induced by a field along this plane. Rather subtle inhomogeneities at specific other atoms for small stabilization and destabilization effects are present. In conclusion, there can be clusters with apparently strong correlations, but these can turn out to be a summarization of data into a cluster without a clear dominant mechanism (or a superposition of different barrier decrease mechanisms). For c12 which was discussed in the main text of Section 7.4.1 (*cf.* Fig. 7.6 on p. 191), the mechanism of the barrier decrease along the positive *z*-plane ESP difference is accomplished by showing a slightly bigger contribution of TSS, contrary to the subtleties and superposed noise in Fig. A.15. For c3, there were also no dominant other correlations except for the one along this plane in order to understand the barrier decrease.

Apart from that, more correlations for the *adaptive* GOCATs discussed in Section 7.4.2 are shown in Figs. A.16 and A.17 on p. 280 and on p. 281.

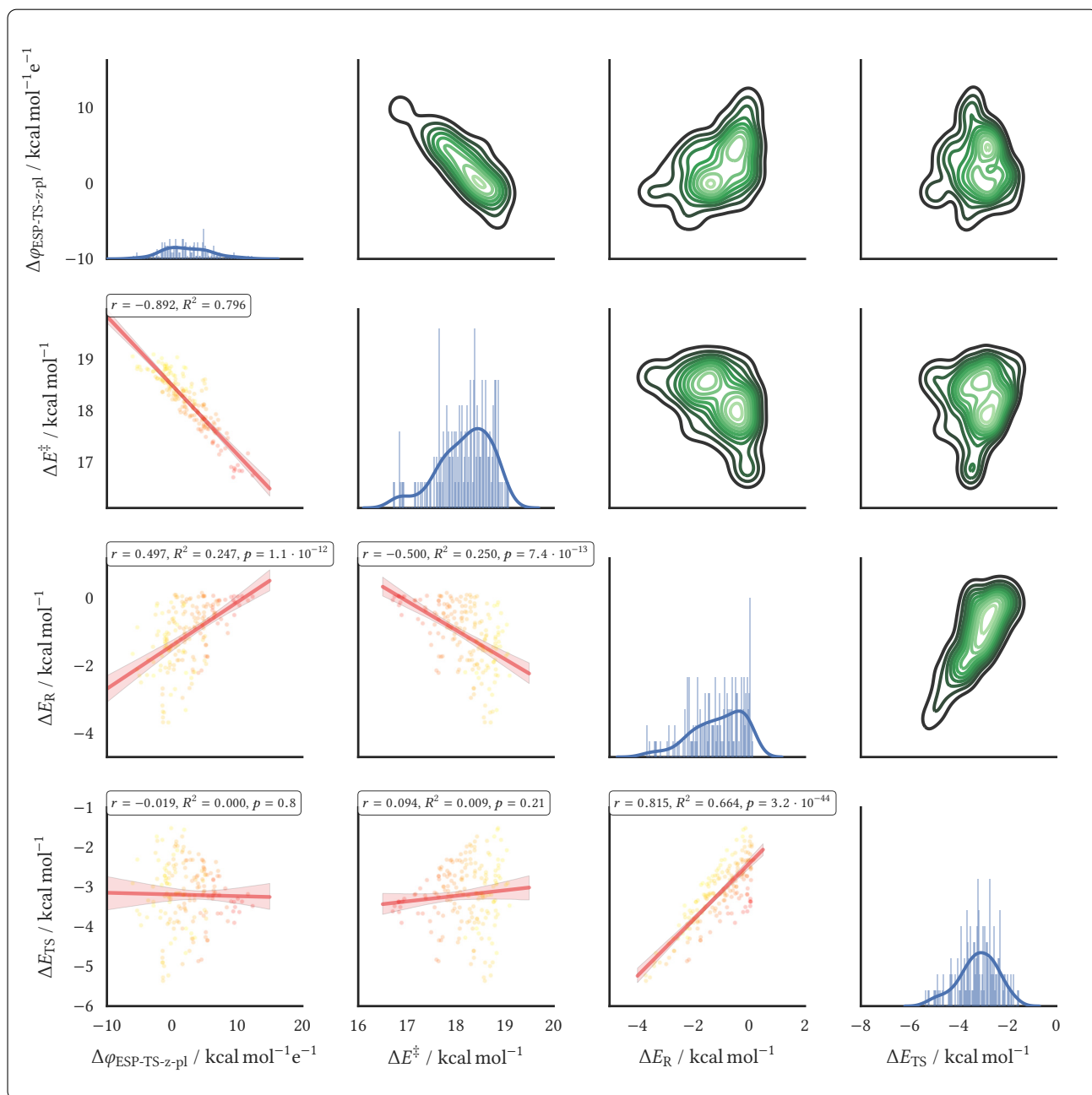


Fig. A.15: $N_{\text{Ch}} = 81$ (sphere, static): Pairwise relationships of the $N_{\text{GOCAT}} = 181$ GOCATs of the highest correlated cluster with respect to the z-plane, c3. For plotting details and for comparisons, see Fig. 7.5. Reaction energy profiles for c3 are shown in Fig. 7.3.

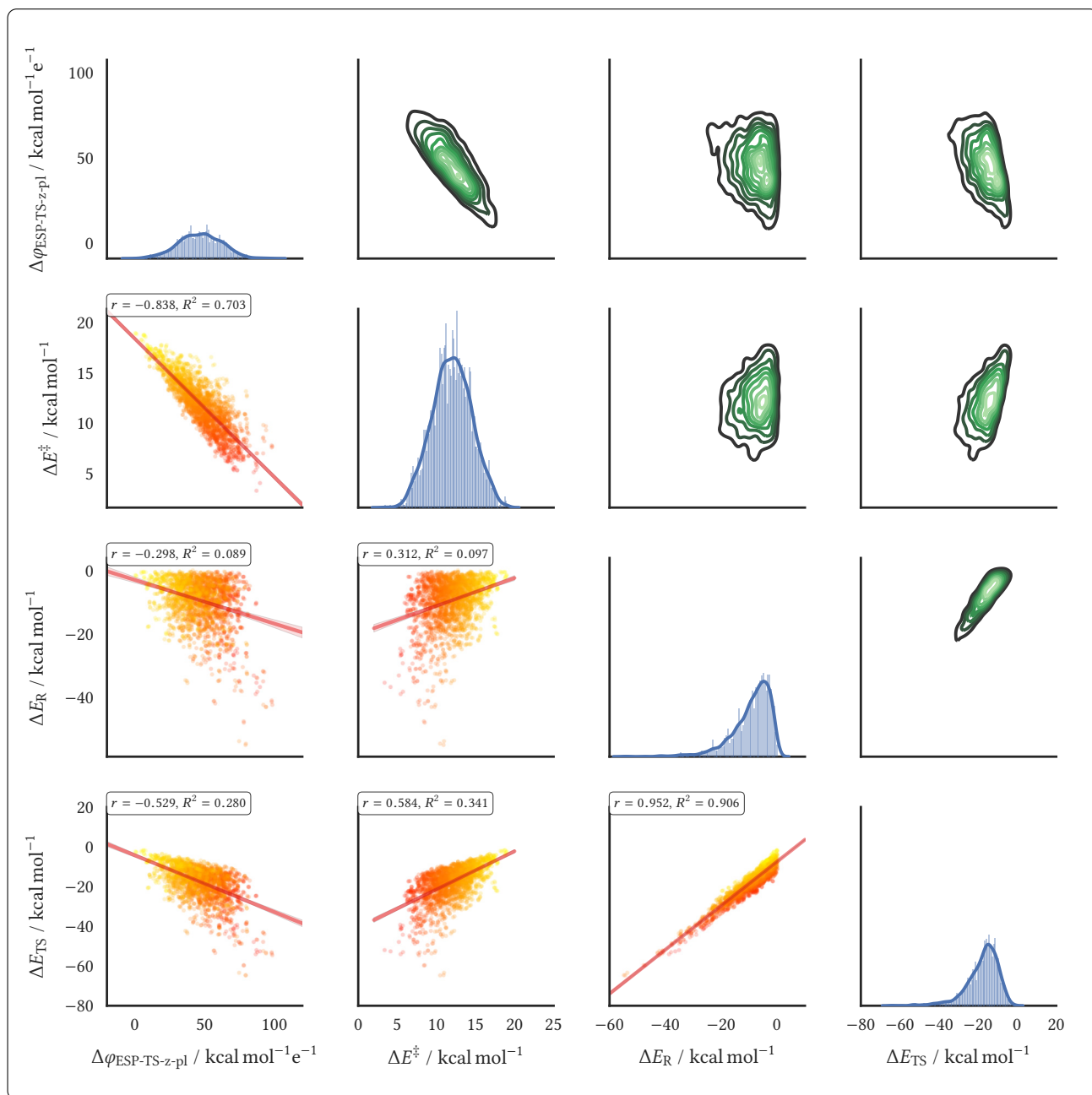


Fig. A.16: $N_{\text{Ch}} = 81$ (sphere, adaptive): Pairwise relationships of the $N_{\text{GOCAT}} = 2952$ GOCATs, complementary to Fig. A.17 on the next page. For plotting details, see Fig. 7.5.

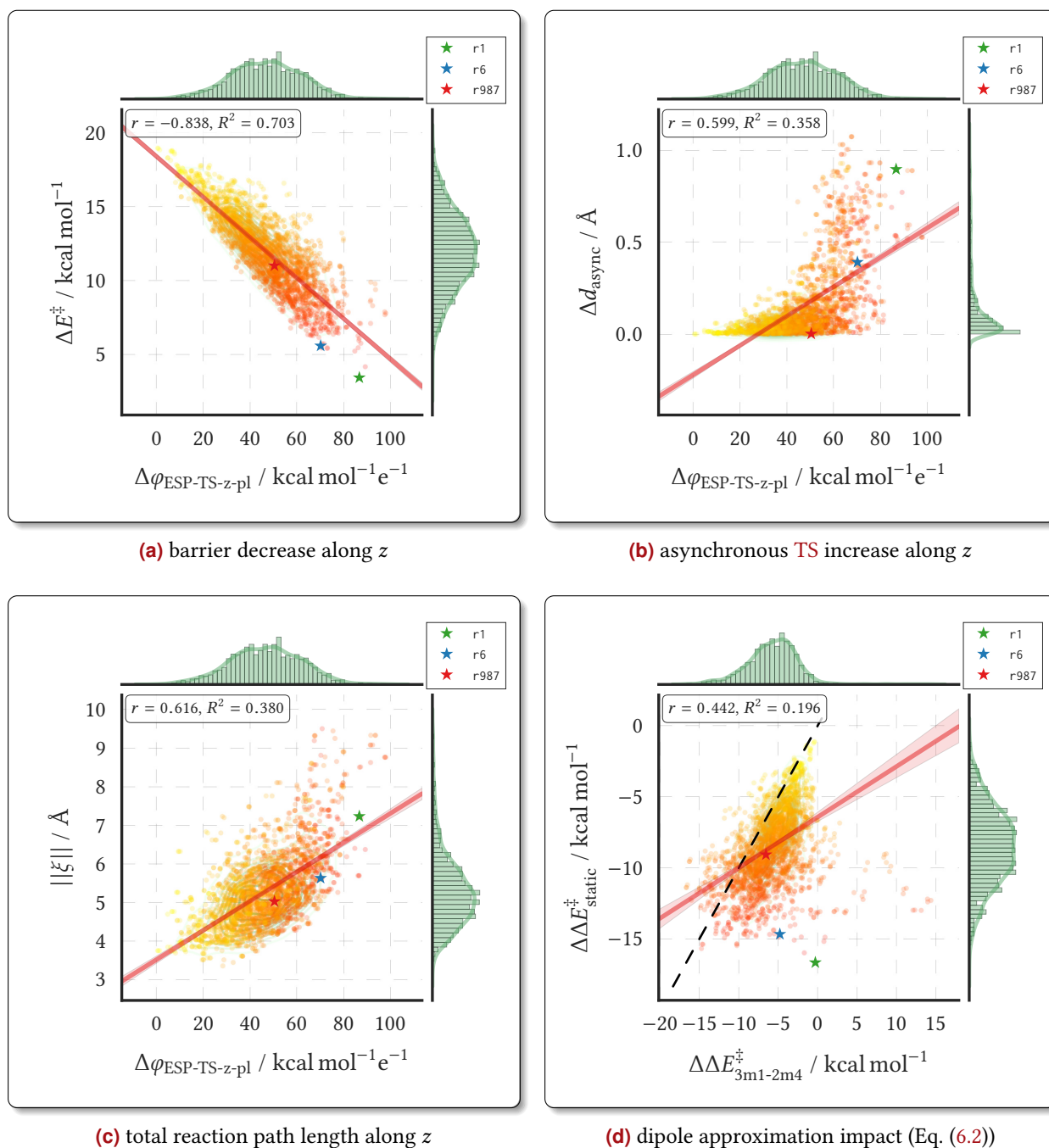
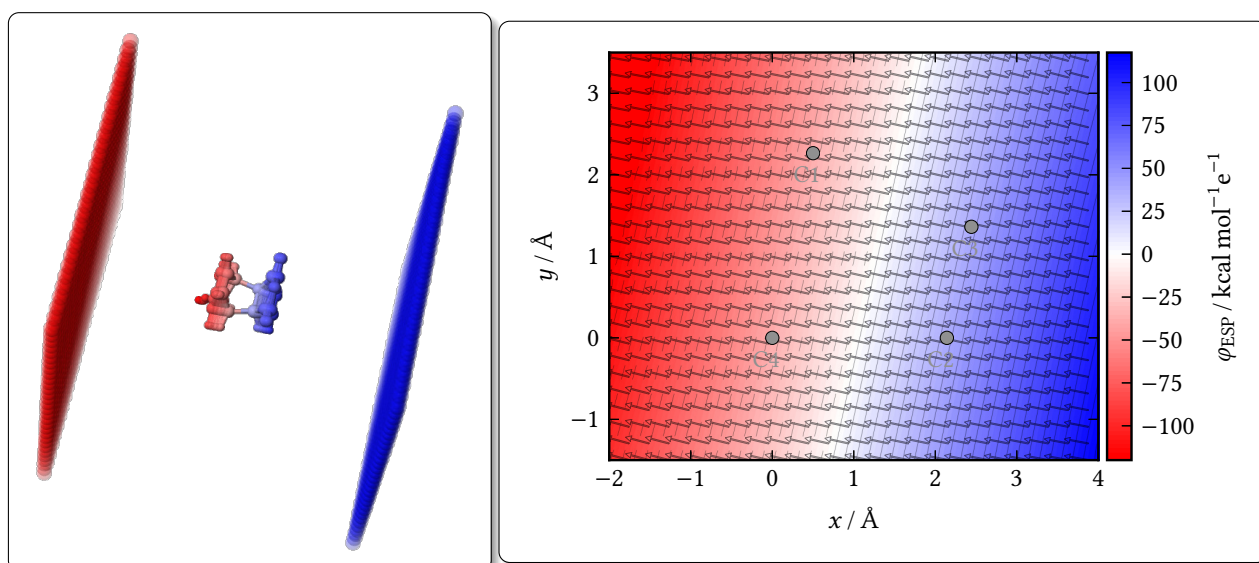


Fig. A.17: $N_{\text{Ch}} = 81$ (sphere, adaptive): Correlations of $N_{\text{GOCAT}} = 2952$ adaptive GOCATs with mean $\|\nabla E\| < 5 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$. Color map between red/yellow for low/high fitness values with the three selected GOCATs ($r1$, $r6$, $r987$) marked explicitly, complementary to Section 7.4.2. Besides the properties discussed in Section 7.4.2, $\|\xi\|$ and Δd_{async} denote the summed reaction coordinate of all frames and the difference between the two new CC-bonds (to be created) at the TS frame, respectively, as proxies for the asymmetry and mechanistic change. A simple linear regression is plotted with a red line, including 95% confidence intervals, and corresponding histograms of the data points are given in the margins, including kernel density estimates, which are also present in the main graphic in the background.

A.5 Uniform Electric Field Data for the Diels–Alder Reaction

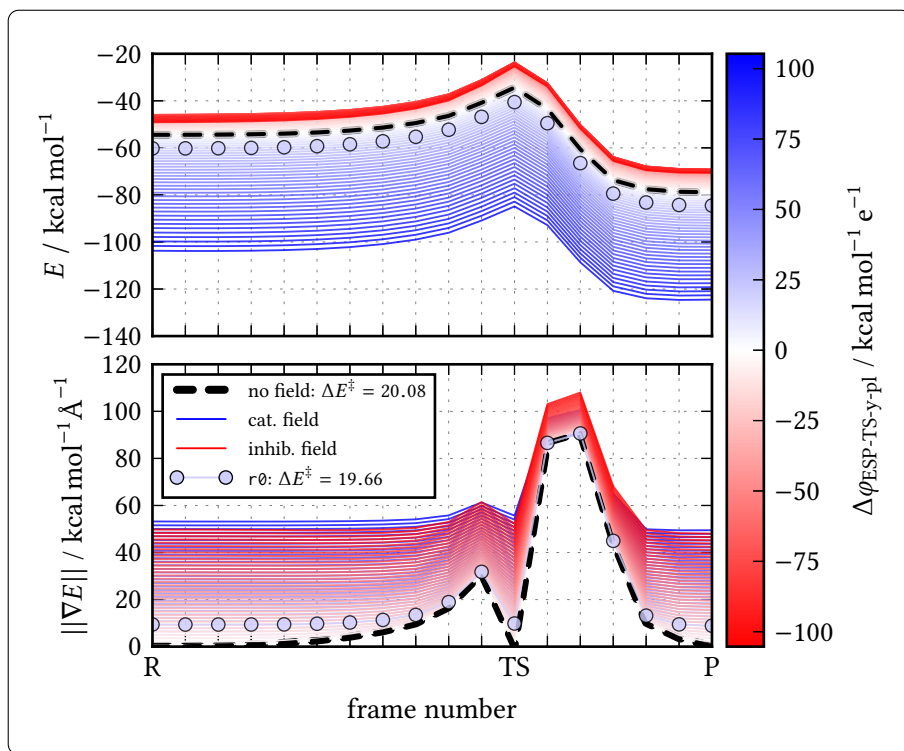
Some more data for the uniform plate GOCATs is given in this Section which is complementary to the discussions in Section 7.4.3. An example of the “plate capacitor” GOCATs is shown in Fig. A.18. Reaction energy profiles for the other directions (y and x) for the *endo* DA reaction follow in Fig. A.19 on the following page. Similarly, the *exo* case is given in Fig. A.20 on p. 284. Correlations and reaction energy profiles for the uniform plate GOCATs are illustrated in Fig. A.21 on p. 285. Finally, reaction energy profiles of adaptively re-optimized MEPs in these plate capacitor GOCATs are shown in Fig. A.21 on p. 285.



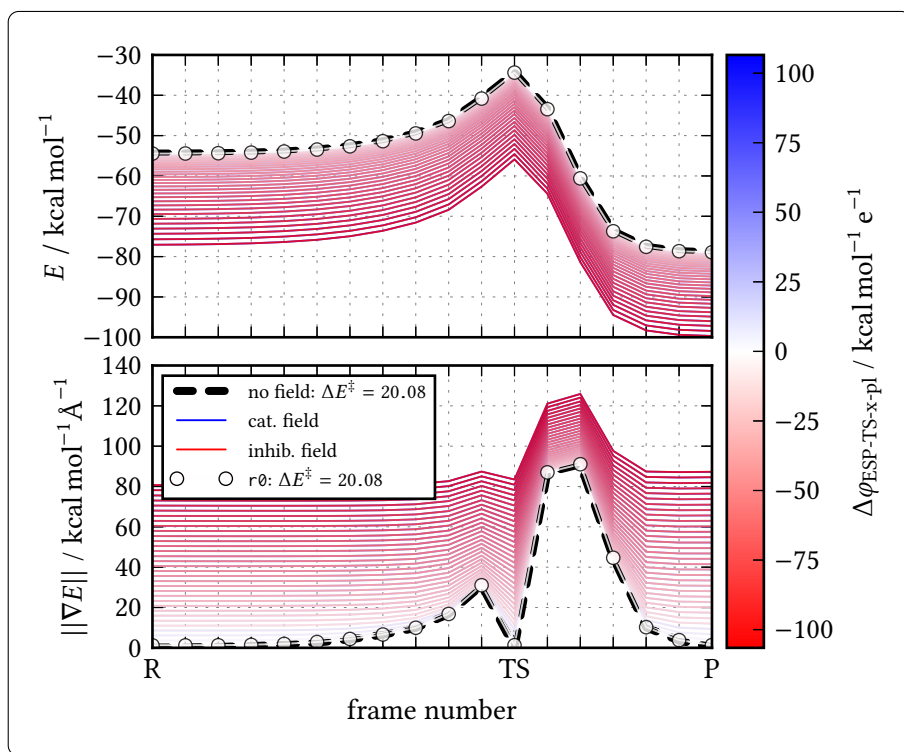
(a) illustration of a uniform GOCAT

(b) ESP and EF at the TS

Fig. A.18: An example for a “plate capacitor” GOCAT is shown in Fig. (a). In Fig. (b), the ESP, ϕ_{ESP} , and its EF as the negative gradient plotted as arrows, are illustrated, projected onto the plane of the four atoms (C1–4) for the TS frame of the *endo* DA reaction shown for this baseline approach. The simple uniform z -field is evident which is emulated by the “plate capacitor”. For plotting details, see Fig. 7.9.

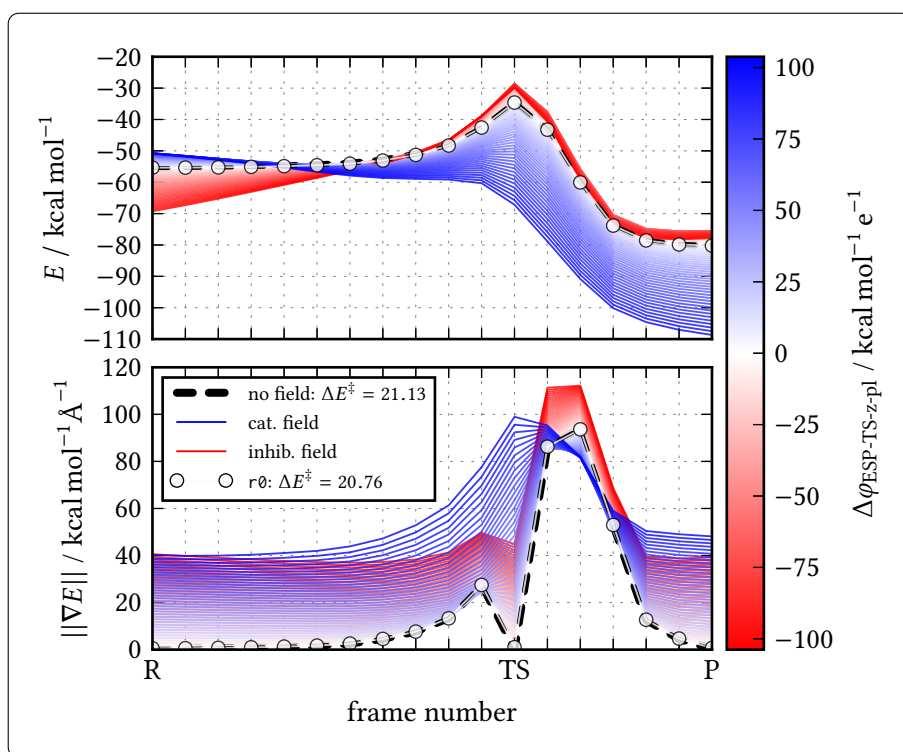


(a) field in y -direction

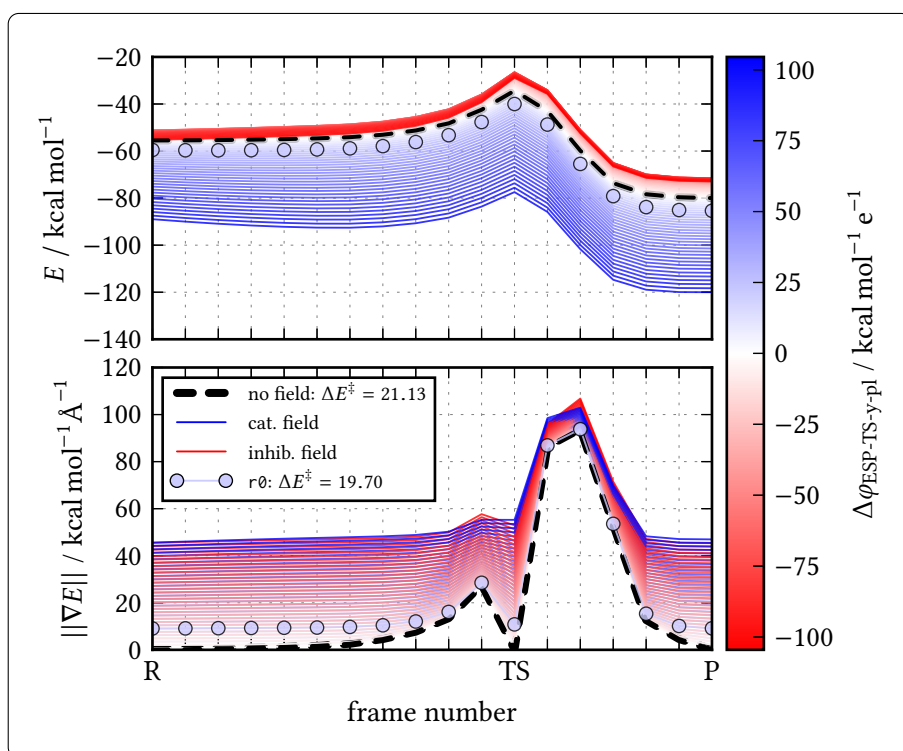


(b) field in x -direction

Fig. A.19: Plate *GOCATs* (static, *endo*): Further reaction energy profiles in uniform electric fields complementary to Fig. 7.15 (p. 201). Compare with Fig. 7.1 for the axes definitions. Note that in Fig. (b) the energies of both positive and negative fields are exactly superposed as expected due to the symmetry of the *DA* structures.

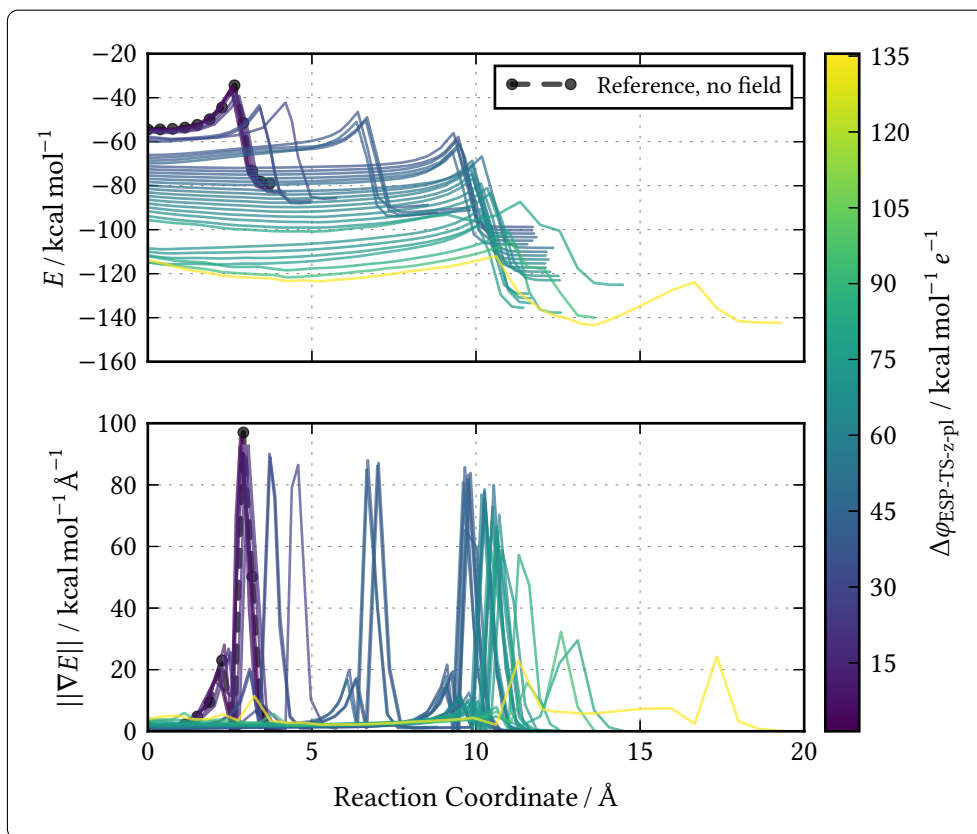


(a) field in z -direction

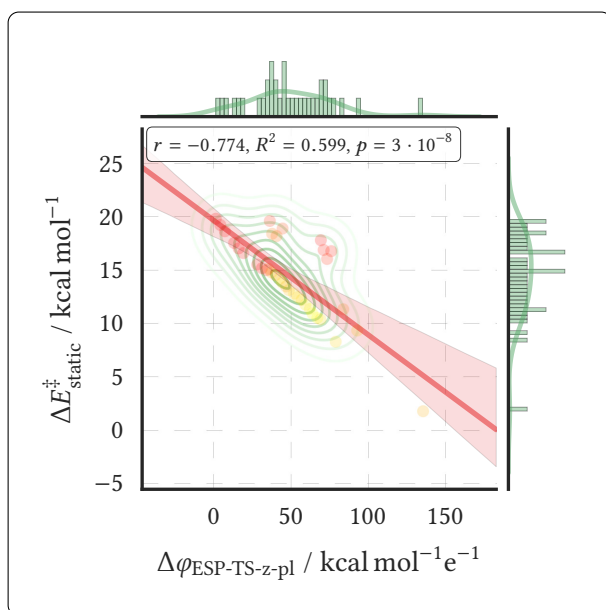


(b) field in y -direction

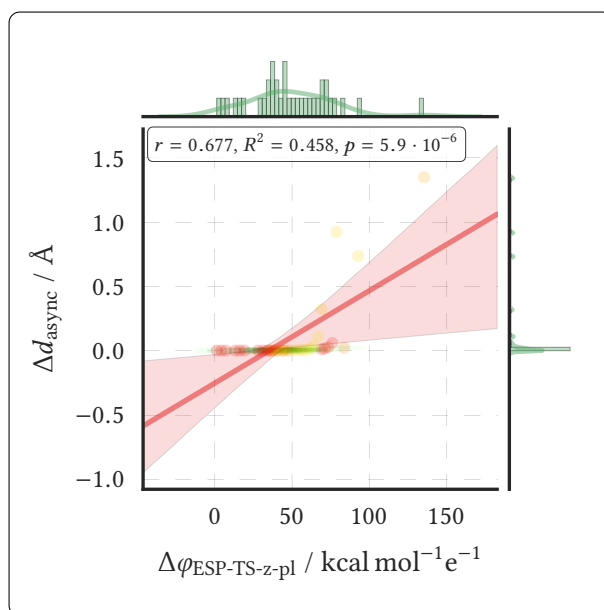
Fig. A.20: Plate GOCATs (static, *exo*): Reaction energy profiles in uniform electric fields complementary to Fig. 7.15 (p. 201), but for the *exo* DA reaction. Compare with Fig. 7.1 for the axes definitions. No x -field reaction energy profiles are shown for *exo* anywhere as these are equivalent to Fig. A.19(b), *i.e.*, no barrier effect at all, but some fully symmetrical shifts in both field directions.



(a) reaction energy profiles



(b) barrier decrease along z



(c) asynchronous TS increase along z

Fig. A.21: Uniform plate GOCATs (adaptive): Fig. (a): Reaction energy profiles after relaxation via Algorithm 3.2. Note that most candidate solutions have $f \gg f_{\text{ref}} = 220.87$ (over-stabilization between R and TS because of convergence issues during the local end-frame and NEB optimizations). In Figs. (b)-(c): Correlations for $\Delta E_{\text{static}}^{\ddagger}$ (barrier always between highest energy and the first frame) and Δd_{async} (difference of CC-distances) with respect to the ESP difference along z . Complementary to Figs. 7.10 and A.17. See also Fig. A.17 for plotting details.

Supplementary Information for the Publications

For convenience and based on a request of one of the referees, the ESI documents of the publications of Chapters 4 to 6 are reprinted in the following. If vectorized versions of the included figures are favored, the reader is referred to the original supplementary documents of the publications.

supplementing Section 4.2	M. DITTNER <i>et al.</i> , EFFICIENT GLOBAL OPTIMIZATION OF REACTIVE FORCE-FIELD PARAMETERS, <i>J. Comput. Chem.</i> 2015 , <i>36</i> , 1550–1561, DOI: 10.1002/jcc.23966. ^[244]	<i>cf.</i> pages 288–293
supplementing Section 5.2	M. DITTNER <i>et al.</i> , CONQUERING THE HARD CASES OF LENNARD-JONES CLUSTERS WITH SIMPLE RECIPES, <i>Comput. Theor. Chem.</i> 2017 , <i>1107</i> , 7–13, DOI: 10.1016/j.comptc.2016.09.032. ^[449]	<i>cf.</i> pages 294–298
supplementing Section 6.2	M. DITTNER <i>et al.</i> , GLOBALLY OPTIMAL CATALYTIC FIELDS – INVERSE DESIGN OF ABSTRACT EMBEDDINGS FOR MAXIMUM REACTION RATE ACCELERATION, <i>J. Chem. Theory Comput.</i> 2018 , <i>14</i> , 3547–3564, DOI: 10.1021/acs.jctc.8b00151. ^[433]	<i>cf.</i> pages 299–362

Supplementary Information for:

Efficient global optimization of reactive force-field parameters

Mark Dittner^(a), Julian Müller^(a), Hasan Metin Aktulga^(b,c) and Bernd Hartke^{(a)}*

(a) Institute for Physical Chemistry,
Christian-Albrechts-University,
Olshausenstr. 40,
24098 Kiel, GERMANY

(b) Dept. Computer Science and Engineering
Michigan State University
428 S. Shaw Lane, Room 3115
East Lansing, MI 48824, USA

(c) Computational Research Division
Lawrence Berkeley National Laboratory
1 Cyclotron Rd, 50F-MS 1650
Berkeley, CA 94720, USA

* corresponding author, hartke@pctc.uni-kiel.de

*J. Comput. Chem., submitted March 12, 2015;
revised version submitted May 8, 2015*

For readers who are not yet familiar with the usual training set definition in ReaxFF fitting, we give here some additional information. For complete compatibility to older more or less *local* optimization algorithms that were used before (e.g., in the van Duin group), we have also implemented the same format and general framework. The following section gives some underlying background of point A of the paper: “definition of the optimization problem”.

General considerations:

For the purpose of fitting a force field to reference data, in principle it would be sufficient to minimize energy deviations at a suitably chosen set of points (geometries). In practice, however, and from the viewpoint of chemistry, it turns out to be more useful, efficient and transparent to also include comparisons of other items: Gradients and frequencies encode information from small regions of the potential energy surface (PES). Results from local geometry optimizations (with given starting points) implicitly contain information on a whole “downhill” path on the PES. And items like charges allow to probe only certain contributions to the total energy (both at the force-field level and at the reference level).

Therefore, typical training sets consist of a mixture of items of these and similar types. Nevertheless, in our framework of general supervised learning non-linear regression via Genetic Algorithms, we stick to the *single-objective* scheme by using as objective function $f(\boldsymbol{\theta})$ one single *weighted sum* (“error sum”) aggregating all the diverse training set items:

$$f(\boldsymbol{\theta}) = \sum_{i=1}^n \left(\frac{y_{\text{ref},i} - y_{\text{calc},i}(\boldsymbol{\theta})}{\sigma_i} \right)^2. \quad (1)$$

Here, the vector $\boldsymbol{\theta}$ contains the parameters to be optimized, and each summation term contains a reference value, $y_{\text{ref},i}$, the corresponding ReaxFF value calculated with the current parameter set $\boldsymbol{\theta}$, and a weight σ_i . For simplicity, this objective function was directly equated to the “fitness” in the main text (although in Genetic Algorithms typically an additional mapping is employed, to generate the fitness from the objective function). σ as a weight fulfills different purposes here: 1.) It allows to include *different* expected standard deviations for *different* properties (e.g., an energy vs. a bond-length); 2.) it scales these contributions to comparable magnitudes; 3.) it makes the overall expression dimensionless, despite the presence of very different observables; 4.) it allows to introduce further prior knowledge, e.g., by different relative weights for whole blocks of different properties (e.g., charges vs. energies) or for differences in chemical importance (properties at minima or transition states vs. those at higher-energy regions).

Very similar definitions were also used in all single-objective ReaxFF optimizations we know of. For a general comparative discussion of this scheme in contrast to the multi-objective approach see the main text of the paper.

At the moment, our OGOLEM setup allows for a broad array of properties y in Eq. 1: energies (absolute and difference energies, optionally scaled for describing “reaction

equations”), geometries (bond lengths, bond angles, dihedral angles), gradient vectors, gradient norms, heat of formations, cell parameters, partial charges, molecular dipoles and excitations energies. In each case, there is a choice of evaluating the property directly at the given molecular geometry (“single point”) or after performing an additional local geometry optimization, with the current force-field parameters θ .

Of course, the search space itself must be also defined beforehand. Thus, from all possible ReaxFF parameters for all chemical elements present in the training set, the subset of parameters, θ , that are to be optimized must be specified, which fixes the dimensionality of the search space. Additionally, for each of these parameters, lower and upper numerical limits have to be chosen (fixing the extent in each dimension). While setting such limits may not be necessary in principle, it makes the definition of many GA operators simpler, it allows to exclude known regions of parameter values in which the ReaxFF equations lose their intended physical meaning, and it simply serves to keep the search space small. In contrast to the procedures used in the van Duin group, we do not need to pre-select any parameter increments.

After this, the actual “optimization of force-field parameters” (point B in the parameter optimization task subdivision explained in the introduction of the main text) can start, which is shown for two pre-definitions of A in the paper. With the highly efficient *global* algorithm we use, there are no significant biases from starting points for the parameters, hence we use random starting points within the parameter boundaries. Also, with the advances reported in the main text, overall real-times for global parameter optimizations are short enough to allow for batches of many long runs in practice. Hence, we can have high confidence in reaching very good minima and quite likely also the best (global) minimum (point B). Therefore, users of this parameter optimization approach can now ignore former limitations in point B and instead focus on iterating and improving points A and C, i.e., on testing resulting optimized force fields and changing the optimization problem definition/setup, until satisfaction, for their application task.

For item C, it is important to point out that due to the big population of current solutions in *one* GA optimization run in B and our *niching* techniques, there still is considerable diversity at the end of convergence, i.e., a *set* of optimized force fields, not just a single one, representing different good local minima in search space. To our experience, several of them should be tested subsequently, e.g., in molecular dynamics, unless the training set setup already is perfect.

Some specifics of the RSSR case:

The choice and design of the training set for the RSSR problem in this publication followed two central considerations:

- The training set should be relatively small and balanced. The size of the training set allows to finish a big number of calculations within reasonable time. Balanced means that the times used for the two main tasks of the fitness evaluation, namely energy calculations and geometry optimizations, are approximately equal.

- The training set data should be assembled in a fashion that would allow this process to be automated. This is because the input set used stems from a line of training sets that were build to check the possibility of blackboxing the ReaxFF parametrization.

We would like to point out that these RSSR training sets are under continuous development in our group, as part of an application project. Therefore, the actual RSSR training set used in the present work was already improved upon between initial submission and revision of the manuscript. Therefore, we refrain from presenting and discussing all details, since some of them are obsolete and full information will be provided in an upcoming publication. Nevertheless, for illustration purposes, we provide some comments on how the RSSR training set used in this work was constructed.

Energies

The reference points for the energy data were taken from thermal trajectories at different finite temperatures, propagated for 5 ps each. The trajectories were calculated at the PM6 level of theory and at temperatures of 100 K, 200 K, 300 K, 400 K and 500 K. To get a representative random sampling from most of the important regions of the potential energy surface, 500 structures were taken from these trajectories. This procedure should additionally enforce a weighted sampling that favors the parts of the phase space that are passed most often by the system, since structures from those parts are present more often in the random pool of structures. This “ad hoc” weighting should lead to a better parametrization of those more important regions. These structures are then prepared as input for single-point calculations with the RIMP2/aug-cc-pVDZ method to get an accurate estimation of their energetics.

Since the energies parametrized in the global fitting routine are relative energies, a locally relaxed structure was prepared and its energy was used to define the zero of energy, arbitrarily but for both the reference level and the force-field level. This geometry optimization was done using the RIMP2/aug-cc-pVDZ method, too. The local optimum structure and the procedure used is schematically shown in Fig. 1.

Due to convergence issues in the SCF routine only 487 of the 500 selected structures made it into the training set. The training set entry for an energy item is of the following general form:

```
ENERGY
1.00 + 300K-247/1 - base/1          313.7
ENDENERGY
```

The actual entry is enclosed in the keywords “ENERGY” and “ENDENERGY” which mark start and stop of the energy input block within the training set file. The first number of the entry is its weight in the accumulated sum of errors. It is followed by two blocks, each containing a sign, an identifier and a divisor. In this case, this means: The energy of structure 300K-247 divided by one is added to the energy, after that the energy of `base` divided by one is subtracted from that value. The resulting energy should be 313.7 kcal/mol, which is the ab-initio reference value. The weight for all energy entries

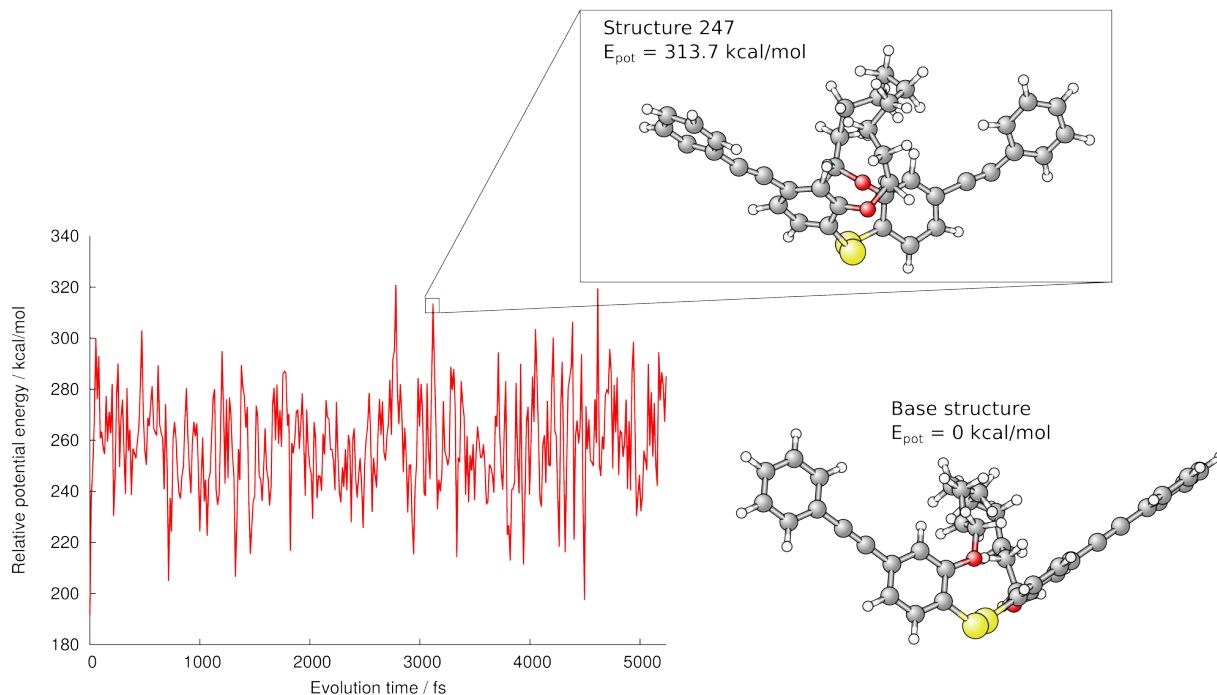


Figure 1: Example for the choice of a structure from a trajectory at 300 K. For the parametrization, the difference of the “Structure 247” energy and the relaxed energy (“Base structure”) is calculated and inserted into the training set.

was chosen to be one, all other weights in the training set were set to be in proportion to this normalized weight of one.

Geometries and charges

The pool of geometries consists of derivatives of parts of the mother compound as found in Figure 1. Since the partial charges of every atom need to be calculated for every energy evaluation and geometry optimization in the ReaxFF formalism, it is convenient to have overlapping structure pools for geometry optimizations and charge calculations in the training set. Figure 2 shows a representative subset of the 32 structures used in the RSSR training set.

The geometries of the reference structures were optimized on the RIMP2/aug-cc-pVTZ level of theory with tight convergence criteria. To stick to the automation paradigm, the initial structures were generated by functionalizing building blocks of the mother compound in the SMILES format. Starting from these optimized geometries the atomic partial charges were calculated using the CHELPG routine and the same ab-initio method as before. CHELPG charges were used because they were found to give partial charge distributions that reproduce experimental values for dipole moments more accurately than other methods.

The inputs for charge and geometry items in the training set differ slightly from the energy items. As seen before, the blocks are enclosed in keywords that mark beginning

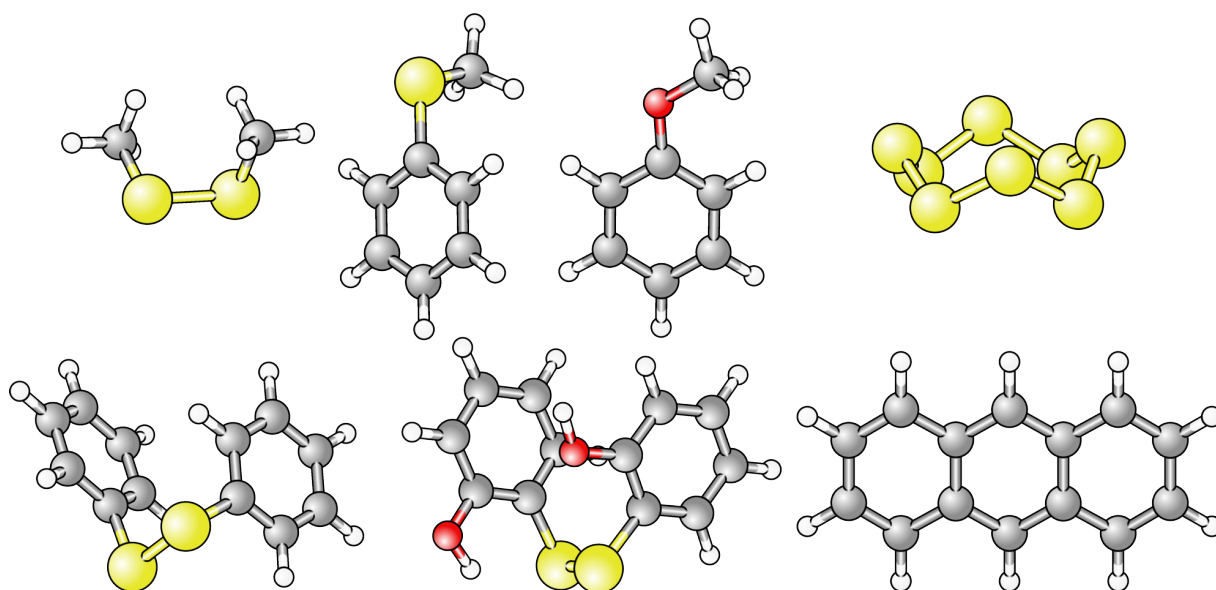


Figure 2: Example structures from the pool of geometries used in the RSSR training set, their short names as used in the training set are given in parenthesis (from top left to bottom right): Dimethyldisulfide (dmds), Thioanisole (tani), Anisole (ani), Sulfur8 (s8), Diphenyldisulfide (dpds), Dihydroxyphenyldisulfide (dpods) and Anthracene (atc).

and end of the respective section. The input line starts with a structure identifier and is then followed by the weighting factor, one to four atom numbers and the ab-initio reference value. If just one atom number is given, the charge of the atom is calculated. Two, three and four atom numbers correspond to bonds, angles and dihedral angles, respectively.

CHARGE

```
cani 0.02 1 -0.480
```

ENDCHARGE

GEOMETRY

```
ani 0.02 2 1 1.388
ani 3.00 1 2 3 120.430
ani 3.00 1 2 3 4 0.026
```

ENDGEOMETRY

In contrast to the weighting factor for the energies, the factors in this example were chosen to be 0.02 for charges and bond lengths and 3.00 for angles. This reflects the different tolerable deviations in these values. Since an energy deviation of 1 kcal/mol is considered to be chemically accurate, a deviation of 1.00 Å in bond lengths would be catastrophic for all applications. The weights used in the RSSR training set should therefore ensure that the molecular properties behave qualitatively correctly when varying in the given ranges.

Supplementary Information for:

Conquering the hard cases of Lennard-Jones clusters with simple recipes

*Mark Dittner and Bernd Hartke**

Institute for Physical Chemistry,
Christian-Albrechts-University,
Olshausenstr. 40,
24098 Kiel, Germany

* corresponding author, hartke@pctc.uni-kiel.de

*Comp. Theor. Chem.,
first version submitted September 1, 2016;
revised version submitted September 20, 2016*

In the following, more detailed results for niching based on nearest-neighbor configurations (NC) and Coulomb matrices (CM) are given in tabular form. *NC mode 1 & 2* as well as *CM setting 1 & 2* are explained in the main text of the article. In all cases, the average global iteration number is given (averaged over 20 runs each), the earliest encounter (minimum number) of iterations needed in the “best” run, the latest encounter (maximum number) in the “worst” run and the standard deviation. Tables 1–5 are plotted in the main article.

All iteration numbers are given as such, not as thousands of steps as in the main article. However, they are again rounded to 2 significant digits. Note that in the other supplementary material (csv files), all tables are provided with raw, unrounded data (6–7 significant digits).

NC-niching

Table 1: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and for $n=100$, for 20 runs using niching *NC mode 1*.

cluster size	average	standard deviation	earliest	latest
75	63000	94000	1100	330000
76	220000	260000	460	850000
77	320000	450000	530	2000000
98	450000	820000	700	2800000
102	32000	74000	1100	250000
103	48000	100000	2500	470000
104	48000	50000	2700	180000
100	24000	30000	3300	98000

Table 2: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and for $n=100$, for 20 runs using niching *NC mode 2*.

cluster size	average	standard deviation	earliest	latest
75	41000	46000	6300	210000
76	80000	100000	6900	360000
77	94000	130000	3900	550000
98	630000	910000	6200	3400000
102	34000	26000	13000	130000
103	24000	14000	5000	66000
104	240000	280000	30000	1200000
100	130000	130000	18000	550000

CM-niching

Table 3: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and for $n=100$, for 20 runs using niching *CM setting 1*.

cluster size	average	standard deviation	earliest	latest
75	180000	290000	23000	1400000
76	290000	400000	15000	1500000
77	900000	1500000	25000	5300000
98	60000	38000	22000	180000
102	50000	16000	25000	95000
103	60000	24000	22000	130000
104	140000	75000	57000	340000
100	40000	14000	21000	77000

Table 4: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and for $n=100$, for 20 runs using niching *CM setting 2*.

cluster size	average	standard deviation	earliest	latest
75	120000	150000	16000	660000
76	320000	350000	22000	1100000
77	310000	580000	27000	2500000
98	220000	470000	30000	2200000
102	53000	16000	26000	91000
103	61000	35000	22000	160000
104	130000	65000	21000	280000
100	57000	19000	29000	92000

Best collection

Table 5: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases, for 20 runs using different nichings.

For each case, the best case that we have observed during our studies is shown (without “meta-optimization” or fine-tunings).

cluster size	average	standard deviation	earliest	latest	niching
75	39000	17000	14000	65000	<i>NC mode 4</i> ^a
76	80000	100000	6900	360000	<i>NC mode 2</i> ^b
77	94000	130000	3900	550000	<i>NC mode 2</i> ^b
98	36000	12000	21000	62000	<i>CM setting 3</i> ^c
102	32000	74000	1000	250000	<i>NC mode 1</i> ^d
103	24000	33000	1200	110000	<i>NC mode 3</i> ^e
104	41000	30000	3500	140000	<i>NC mode 3</i> ^e

^a Equal to *mode 2* (see main article), but using a smaller bin width of 10 percentage points (cf. Table 7).

^b Explained in main text (cf. Table 2).

^c Equal to *setting 1*, but using smaller $d(\mathbf{M}, \mathbf{M}') = 5$ as minimum difference between individuals (atomic units, cf. Table 8).

^d Explained in main text (cf. Table 1).

^e Equal to *mode 1* (see main text), but using a smaller bin width of 10 percentage points (cf. Table 6).

Additional nichings for the collection:

Table 6: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and for $n=100$, for 20 runs using niching *NC mode 3*.

cluster size	average	standard deviation	earliest	latest
75	170000	360000	3500	1700000
76	330000	480000	400	1600000
77	140000	200000	1300	670000
98	1200000	2200000	760	8300000
102	45000	55000	840	180000
103	24000	33000	1200	110000
104	41000	30000	3500	140000
100	24000	47000	3100	210000

Table 7: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and for $n=100$, for 20 runs using niching *NC mode 4*.

cluster size	average	standard deviation	earliest	latest
75	39000	17000	14000	65000
76	130000	190000	13000	740000
77	120000	130000	14000	470000
98	1100000	1700000	12000	6300000
102	39000	17000	14000	79000
103	32000	13000	9000	56000
104	140000	77000	36000	330000
100	92000	70000	23000	250000

Table 8: Number of global optimization steps until first encounter of the true global minimum, for several LJ hard cases and for $n=100$, for 20 runs using niching *CM mode 3*.

Note: Interestingly, this setting with a really small threshold, $d(\mathbf{M}, \mathbf{M}') = 5$, is exceptionally good for $n = 98$ (and $n > 98$ in the LJ sizes studied here), but the worst for $n \leq 75$ (but still better than no niching at all). The reason is discussed in the main text of the article.

cluster size	average	standard deviation	earliest	latest
75 ^a	6600000	5900000	18000	18000000
76 ^b	7100000	5400000	22000	18000000
77 ^c	6800000	4900000	30000	15000000
98	36000	12000	21000	62000
102	60000	84000	25000	410000
103	53000	53000	24000	240000
104	67000	22000	39000	130000

^a For this particular size just 18/20 runs were included in the statistics. The other 2 were not successful at finding the global minimum in $< 20 \cdot 10^6$ global optimization iterations.

^b For this particular size just 15/20 runs were included in the statistics. The other 5 were not successful at finding the global minimum in $< 20 \cdot 10^6$ global optimization iterations.

^c For this particular size just 14/20 runs were included in the statistics. The other 6 were not successful at finding the global minimum in $< 20 \cdot 10^6$ global optimization iterations.

Supporting information for:
**“Globally Optimal Catalytic Fields – Inverse
Design of Abstract Embeddings for Maximum
Reaction Rate Acceleration”**

Mark Dittner and Bernd Hartke*

Institute for Physical Chemistry, Christian-Albrechts-University Kiel, 24098 Kiel, Germany

E-mail: hartke@pctc.uni-kiel.de

Contents

Notes On Software	S3
Overview: Complementary Content	S4
S1 Concerning The Objective Function:	S5
S1.1 GOCAT Gas Phase Optimization	S5
S1.2 GOCAT Stabilizing COSMO Path	S6
S2 PM7: $N_{\text{Ch}} = 1$ case (non-neutral summed charges)	S7
S2.1 Cluster Analysis	S7
S2.2 Reaction Paths	S9
S2.3 Selected Details	S12
S3 PM7: $N_{\text{Ch}} = 3$ case (non-neutral summed charges)	S18

S3.1 Cluster Analysis	S18
S3.2 Reaction Paths (Selected Clusters)	S22
S3.3 Selected Details	S23
S4 PM7: $N_{\text{Ch}} = 10$ case (non-neutral summed charges)	S29
S4.1 Complementary Discussions	S29
S4.2 Cluster Analysis	S33
S4.3 Reaction Paths (Selected Clusters)	S36
S4.4 Selected Details	S37
S5 PM7: $N_{\text{Ch}} = 10$ case (summed charge neutrality)	S39
S5.1 Cluster Analysis	S39
S5.2 Reaction Paths (Selected Clusters)	S41
S5.3 Selected Details	S42
S6 PM7: $N_{\text{Ch}} = 10$ case (summed charge neutrality and no TS gradient norm threshold)	S45
S6.1 Complementary Illustration	S45
S6.2 Cluster Analysis	S46
S6.3 Reaction Paths (Selected Clusters)	S48
S6.4 Selected Details	S49
S7 PM7: $N_{\text{Ch}} = 10$ stabilizing COSMO path (summed charge neutrality)	S51
S7.1 Cluster Analysis	S51
S7.2 Selected Details	S55
S8 DFT: $N_{\text{Ch}} = 10$ case (summed charge neutrality)	S56
S8.1 Cluster Analysis	S56
S8.2 Reaction Paths (Selected Clusters)	S59
S8.3 Selected Details	S60

Notes On Software

As briefly stated in the main text, almost everything is implemented as extension to our (open source) program package `OGOLEM`.^{S1,S2} At the time of writing, the *local* changes for the GOCAT optimization is not yet published anywhere and is an own local fork of the project. For the analysis part, we use corresponding python scripts using open source libraries like `SCIPY`^{S3} and `SCIKIT-LEARN`.^{S4}

So, at the moment, the code can be obtained by request. As this code base is still changing a lot, we have not yet considered to publish it separately or as a merged version with the already downloadable code at <https://www.ogolem.org/>.

Overview: Complementary Content

In the following, we prepared some complementary figures and statistics to the ones in the main article: For the cases $N_{\text{Ch}} = 1$ (Section S2) and $N_{\text{Ch}} = 10$ (Section S4). $N_{\text{Ch}} = 3$ (Section S3) as intermediate case was not shown in the main article. These are the cases with arbitrary (mostly non-neutral) summed total charge of the electrostatic GOCATs. In Section S5, the complementary plots for the neutral GOCATs are given and in Section S7 the ones for the COSMO case. For a glimpse on a higher level of theory, DFT results follow in Section S8. Section S6 was added (on request) and uses the same setting as Section S5, but no TS gradient norm thresholds (i.e., a slightly different objective function from that described in the main article) for *discrete* shifts of the TS in the GOCAT. In the next Section S1, some explicit weights of the used objective functions are given.

S1 Concerning The Objective Function:

S1.1 GOCAT Gas Phase Optimization

More details on the weights used in the typical objective function (Section 2, main article, the following enumeration follows the items there; internal units are atomic ones (a.u.)):

- item 1. and item 3.: $\Delta E^\ddagger = E_{\text{TS}} - E_{\text{E}}$. As intermediate GOCATs can have very curvy paths (multiple local minima/maxima), we used: 10 times *all* barriers found plus 1 time the biggest barrier. At the end (with gradient thresholds below), there is just one barrier left (unimodal). (Otherwise, with too many non-continuous penalties the search space gets too rugged – or just reading out fixed 2-point differences there might be some additional barriers created, overseen otherwise).
- item 2.: Stabilization penalty: $f(\Delta E_{\text{TS,stab.}}) = p \cdot \Delta E_{\text{TS,stab.}}^2$, if $\Delta E_{\text{TS,stab.}} > 0$, usually with $p \approx 42 \cdot 10^3$. Note: At the end of the GA optimization, every single individual has a contribution of 0, as all GOCATs are stabilized.
- item 4.: Here a non-continuous penalty of 1.5 is added, if the read-out minima energies and maximal energy (TS) are 2 frames off the gas phase reference (at the end of GA optimization, zero contribution throughout).
- item 5.: for all $\|\nabla E_{\{\text{E,TS,P}\}}\| > 10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, squared penalty function like in item 2., but using $p \approx 4200$ (i.e. physical dimension differ from energy, but the final fitness is of course dimensionless such that any weight can be used that creates a meaningful balance between gradients and barrier).

At the end (for visualization), the fitness values are scaled by 627.51 (from a.u. to kcal mol^{-1}).¹ Final, typical contributions for e.g.: best GOCAT, **r0**, of size $N_{\text{Ch}} = 10$ (Section S5): $152.39 = 11.92 + 140.47 + 0 + 0$: gradients (a mean of the norms of $11.3 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ for

¹ Side note: As the fitness function implementation progressed from being very simple (i.e., just TS energy, nothing else) to include many more ingredients, the conversion to final intuitive energy values was (once) meaningful. Finally, just the *relations* of ingredients matter, the final absolute values do not in our GA implementation.

{E, TS, P}), barrier of 12,77 kcal mol⁻¹, zero fitness for TS stabilization and zero for other penalties. The gas phase Menshutkin reference reaction has a fitness of 334.95 (all GOCATs with less fitness will have a barrier decrease and just a small gradient norm increase over the threshold).

S1.2 GOCAT Stabilizing COSMO Path

The change of the objective function was described in Section 3.4 (main article).

- item 5.: for all $\|\nabla E_{\{E,TS,P\}}\| > 1 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, squared penalty function like above with $p \approx 21 \cdot 10^3$ (both threshold smaller and weight 5 times bigger than in the other objective function).
- item 4.: Same as above.
- New item: For each reference frame, a squared difference function of energies is used of $f(\mathbf{E}_{\text{GOCAT}}) = \sum_i p_i \cdot (E_{\text{ref},i} - E_{\text{GOCAT},i})^2$, with each $p_i \approx 420$ (i.e., the same weight on each frame i).

The other items (TS stabilization and barrier decrease) are not used.

At the end (for visualization), the fitness values are scaled by 627.51 (from a.u. to kcal mol⁻¹). Final, typical contributions for e.g.: best GOCAT, r0, of size $N_{\text{Ch}} = 10$ (Section S7): 2759.92 = 2582.42 + 177.50 + 0: for gradients (a mean of norms of 10.3 kcal mol⁻¹ \AA⁻¹ for {E, TS, P}), mean summed absolute difference to the COSMO reference energies of 2.87 kcal mol⁻¹ and zero fitness for other penalties.

S2 PM7: $N_{\text{Ch}} = 1$ case (non-neutral summed charges)

S2.1 Cluster Analysis

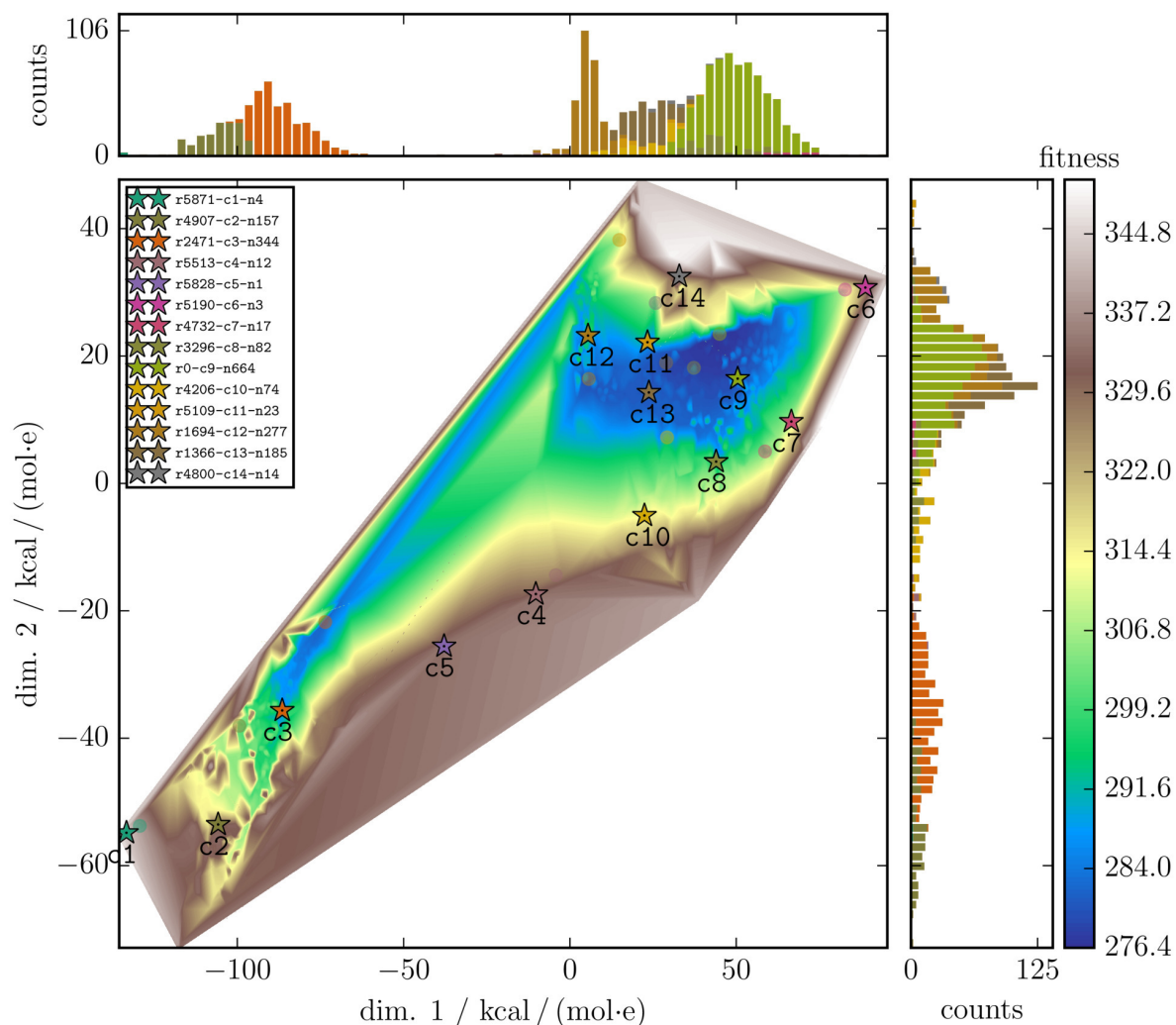


Fig. S1: $N_{\text{Ch}} = 1$ GOCATs: Multidimensional Scaling as 2D projection similar to Fig. 4 (main article). Also a linearly interpolated fitness surface of all individuals in this 2D plot is given (color map). Besides the cluster means (stars) also the best rank of that cluster (the individual indicated by rn in labels of the legend) is plotted as circle.

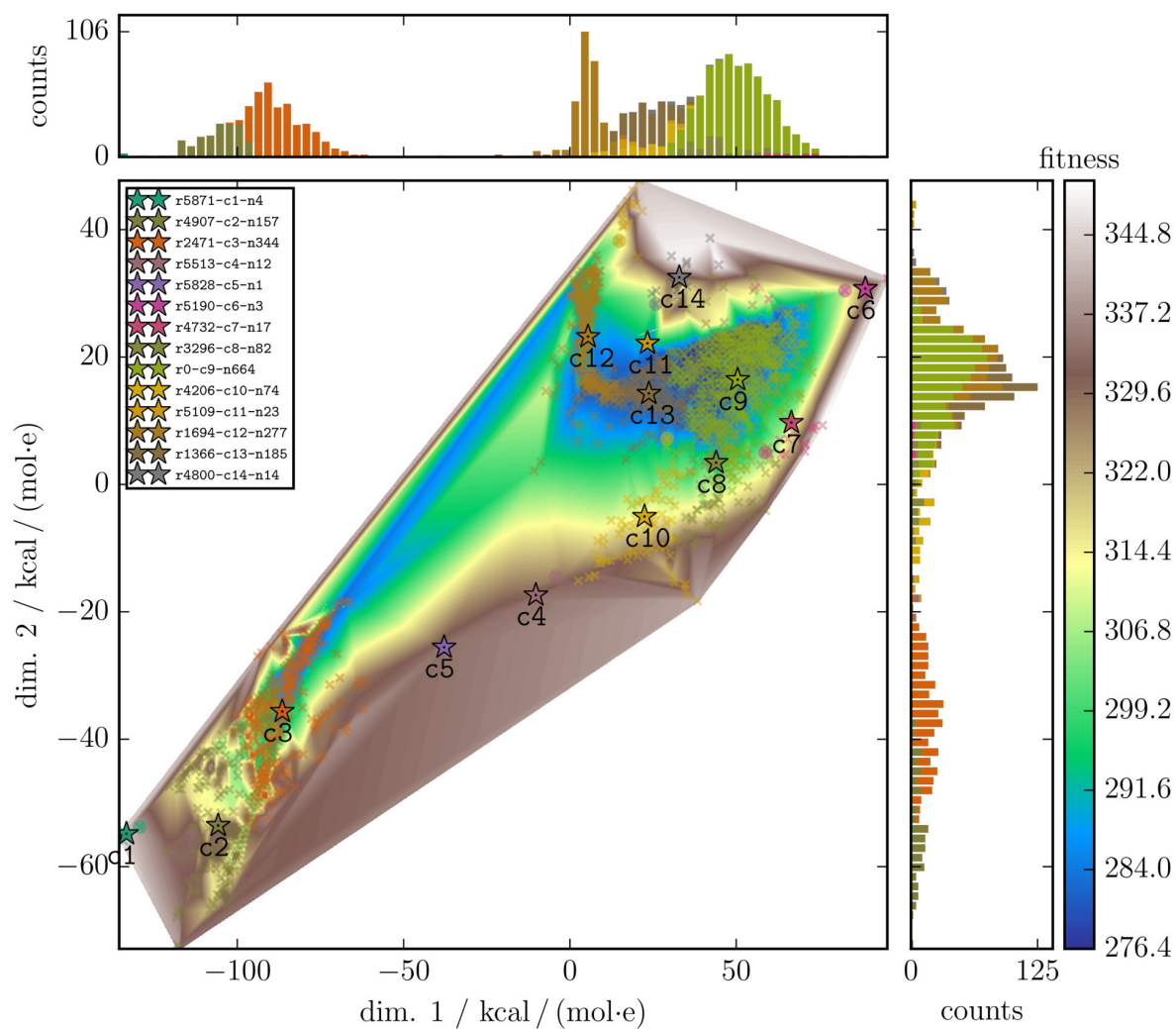


Fig. S2: $N_{\text{Ch}} = 1$ GOCATs: Similar plot to Fig. S1, but showing also the actual points to illustrate dense regions of individuals (and thus regions, where the linear interpolation of the fitness values is well defined) and less dense regions.

S2.2 Reaction Paths

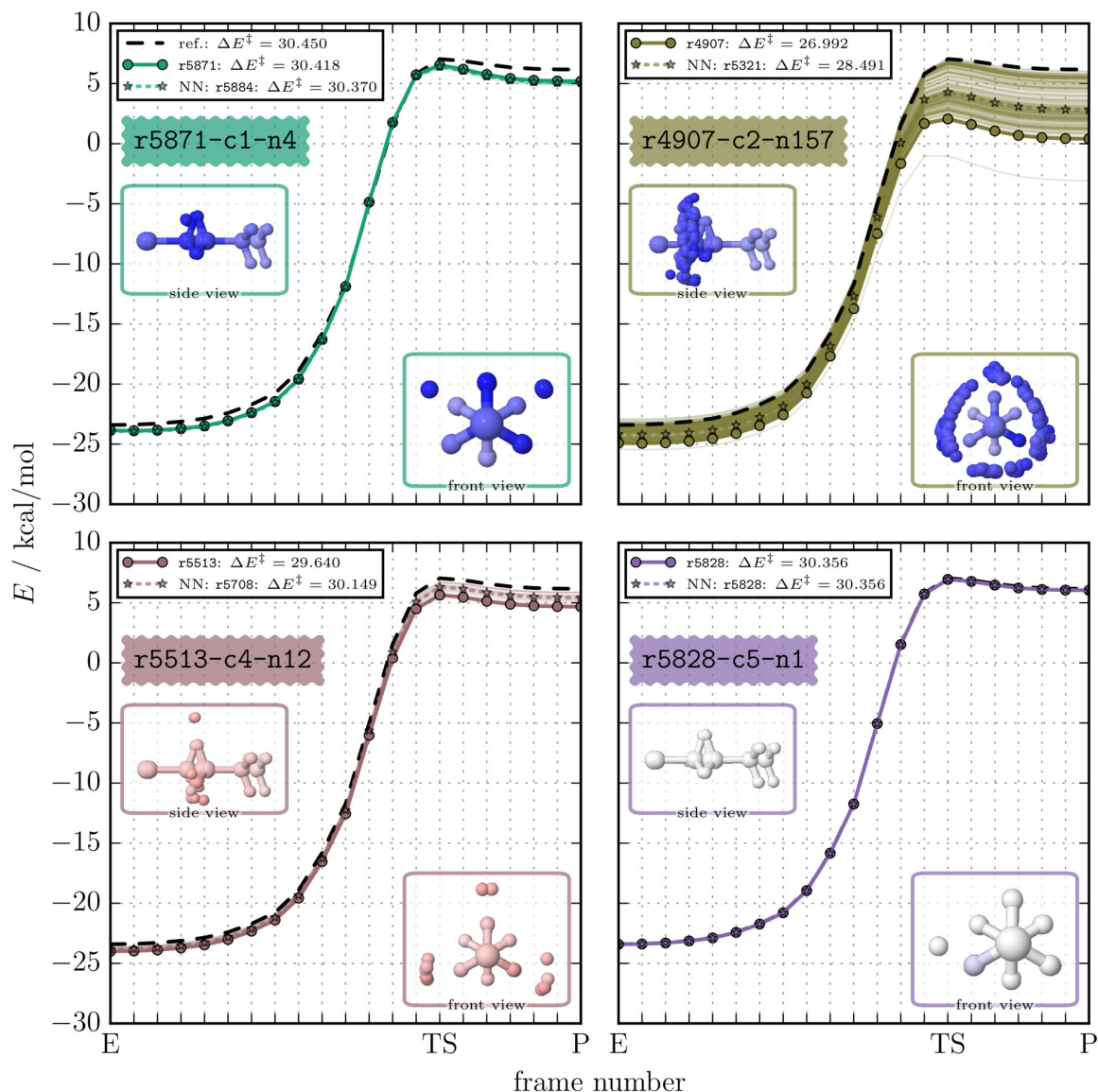


Fig. S3: $N_{\text{Ch}} = 1$ GOCATs (PM7): (Pre-optimized) reaction path for the PM7 reference calculation is shown as well as the path within the electrostatic GOCATs of four different clusters, including also outlier ones (c1, c2, c4, c5). The lowest rank of each cluster (circles) and the nearest neighbor (NN) to the mean ESP vector of that cluster (stars) is plotted separately. All other individuals of the cluster are also plotted with thin lines to illustrate the spread. Two images in different perspectives of the overlain individuals are given in the insets. c5 as corner case almost shows *no* effective GOCAT at all, with barriers and fitness values very close to the filtering thresholds of the raw database before.

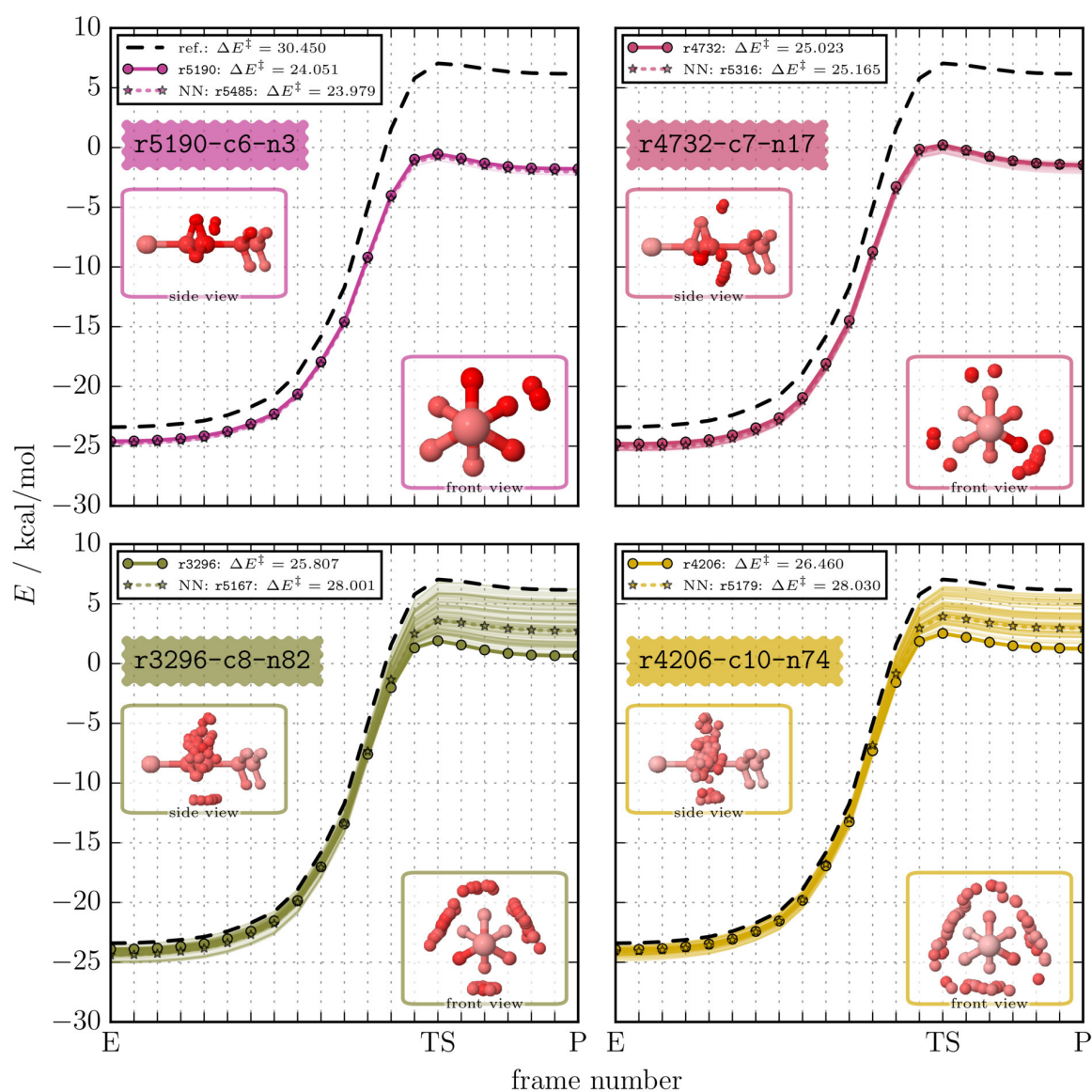


Fig. S4: $N_{\text{Ch}} = 1$ GOCATs (PM7): (Pre-optimized) reaction path for the PM7 reference calculation is shown as well as the path within the electrostatic GOCATs of four different clusters (c6, c7, c8, c10); for illustration details, see Fig. S3.

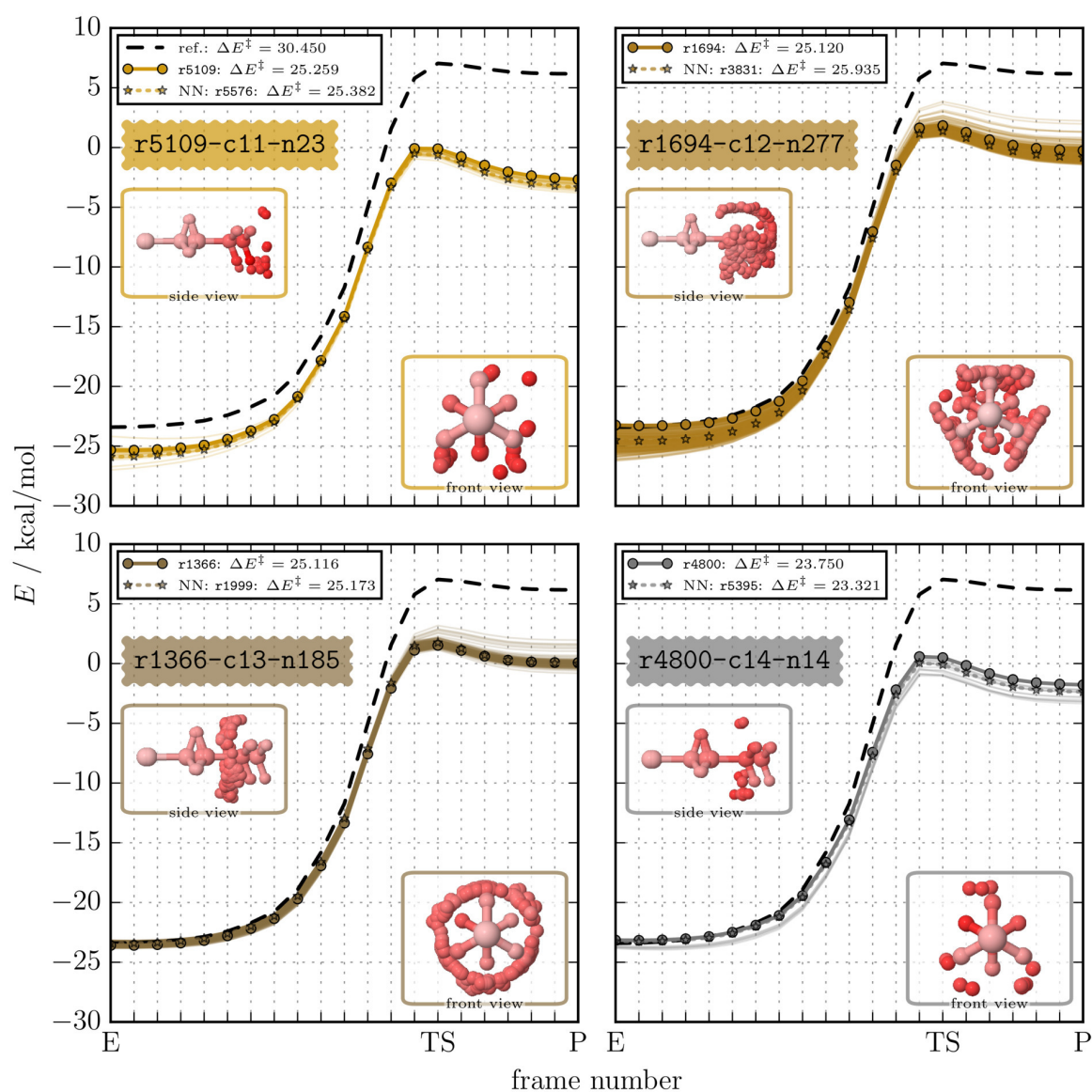


Fig. S5: $N_{\text{Ch}} = 1$ GOCATs (PM7): (Pre-optimized) reaction path for the PM7 reference calculation is shown as well as the path within the electrostatic GOCATs of four different clusters (c11, c12, c13, c14); for illustration details, see Fig. S3.

S2.3 Selected Details

Table S1: Properties of the final clusters for the $N_{\text{Ch}} = 1$ case and of their best rank individual: For clusters, both mean values of those clusters as well as standard deviations (parentheses) are given. For separate individuals, just their single value is presented.

cluster name or individual	fitness	barrier [kcal mol ⁻¹]	grad. norm (E) [kcal mol ⁻¹ Å ⁻¹]	grad. norm (TS) [kcal mol ⁻¹ Å ⁻¹]	grad. norm (P) [kcal mol ⁻¹ Å ⁻¹]	sum. charge (elem. charge)
r5871-c1-n4	343.264(3.423)	30.351(0.069)	11.263(0.321)	10.775(0.332)	11.614(0.320)	0.137(0.004)
r4907-c2-n157	317.469(10.737)	28.642(1.090)	8.033(0.887)	10.310(1.086)	10.233(0.723)	0.109(0.009)
r2471-c3-n344	298.003(15.087)	26.805(1.495)	4.729(1.168)	10.354(1.985)	8.869(1.289)	0.085(0.010)
r5513-c4-n12	331.332(2.492)	30.121(0.227)	2.764(0.819)	3.148(0.960)	3.664(1.071)	-0.036(0.011)
r5828-c5-n1	333.914(0.000)	30.356(0.000)	0.429(0.000)	0.510(0.000)	0.534(0.000)	0.005(0.000)
r5190-c6-n3	325.585(17.295)	23.939(0.136)	11.412(0.607)	13.212(0.363)	14.517(0.664)	-0.166(0.008)
r4732-c7-n17	320.667(15.877)	25.065(0.155)	10.836(0.580)	13.352(0.481)	13.337(0.713)	-0.135(0.007)
r3296-c8-n82	312.639(12.639)	28.382(1.144)	8.142(0.890)	8.176(0.915)	9.701(0.883)	-0.104(0.009)
r0-c9-n664	283.114(6.311)	25.387(0.619)	8.234(0.725)	10.759(0.768)	10.728(0.711)	-0.114(0.012)
r4206-c10-n74	314.904(11.063)	28.625(1.004)	5.951(0.876)	6.339(1.232)	7.387(1.152)	-0.078(0.011)
r5109-c11-n23	327.160(10.357)	25.256(0.569)	12.531(0.738)	14.329(0.594)	10.070(0.628)	-0.123(0.008)
r1694-c12-n277	288.335(6.209)	25.964(0.633)	8.507(1.709)	10.968(0.655)	7.796(0.715)	-0.086(0.013)
r1366-c13-n185	281.215(2.170)	25.231(0.298)	6.356(0.463)	11.246(0.725)	8.639(0.596)	-0.084(0.004)
r4800-c14-n14	326.616(17.802)	23.227(0.288)	8.563(0.362)	15.854(0.724)	11.480(1.091)	-0.108(0.007)
r5871	338.382	30.418	10.793	10.291	11.146	0.131
r4907	300.869	26.992	6.190	11.456	9.744	0.095
r2471	281.757	25.311	5.092	11.336	9.118	0.078
r5513	326.045	29.640	3.450	3.972	4.556	-0.046
r5828	333.914	30.356	0.429	0.510	0.534	0.005
r5190	308.590	24.051	10.799	12.848	13.851	-0.158
r4732	298.541	25.023	9.850	12.804	12.147	-0.123
r3296	283.875	25.807	6.714	9.305	9.178	-0.099
r0	276.588	24.888	7.496	11.176	10.357	-0.110
r4206	291.056	26.460	6.088	8.233	8.089	-0.086
r5109	305.943	25.259	11.450	13.598	9.643	-0.112
r1694	279.870	25.120	7.015	11.379	7.730	-0.071
r1366	278.710	25.116	6.322	11.141	9.366	-0.088
r4800	299.500	23.750	8.077	14.526	9.751	-0.097

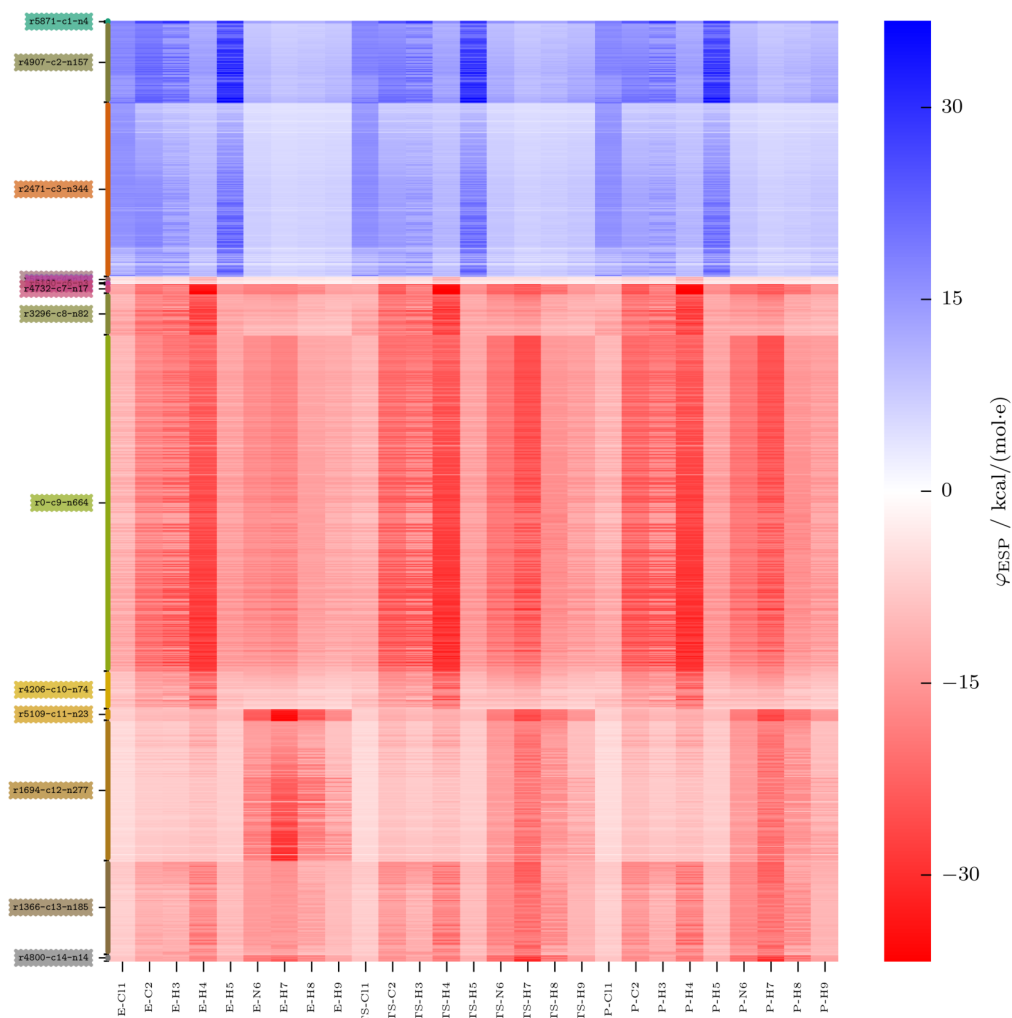


Fig. S6: $N_{\text{Ch}} = 1$ GOCATs (PM7): Heatmap of all explicit electrostatic potential values at the 9 core atoms of 3 selected frames: E, TS and P. The complete database is plotted, chunked into the 14 clusters and each cluster sorted with respect to their rank (lowest rank, i.e., lowest fitness is plotted from top to bottom within each cluster).

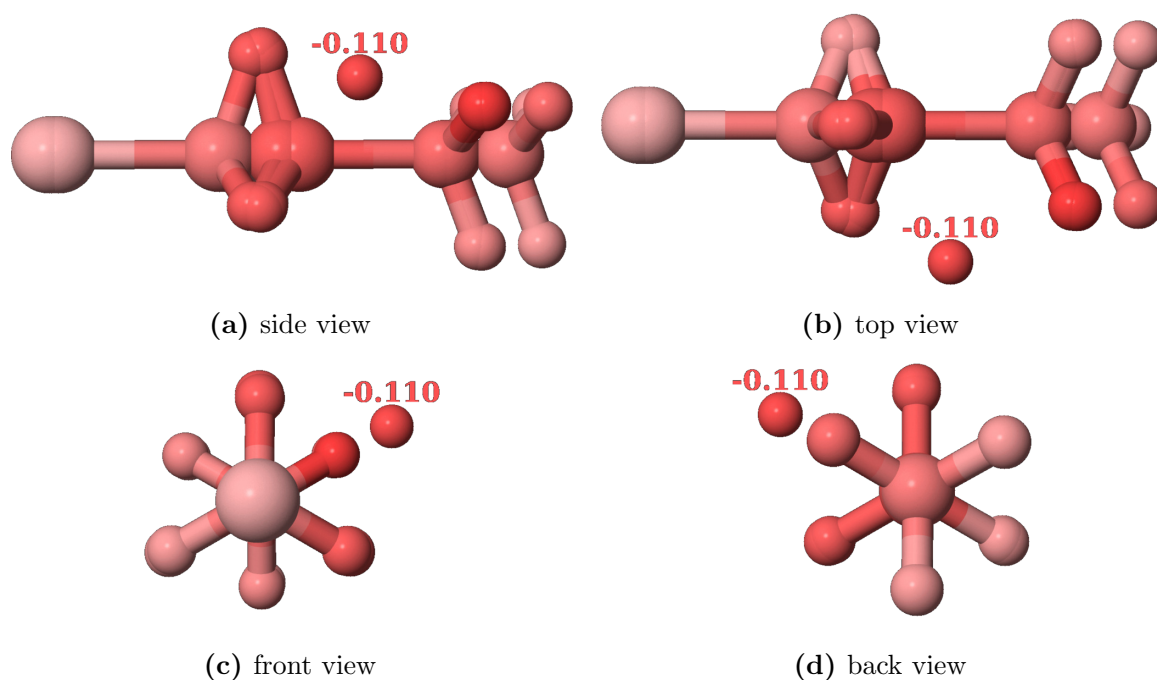


Fig. S7: Single GOCAT individual: four different views of r_0 (of c_9) for the $N_{Ch} = 1$ case with values given for the partial charges. Both partial charges and the 3 · 9 atoms of the selected core frames (E, TS, P) are colored red/blue in the ranges $[-0.174, +0.174] e$ for charges and $[-36.771, +36.771] \text{kcal mol}^{-1} e^{-1}$ for ESP values.

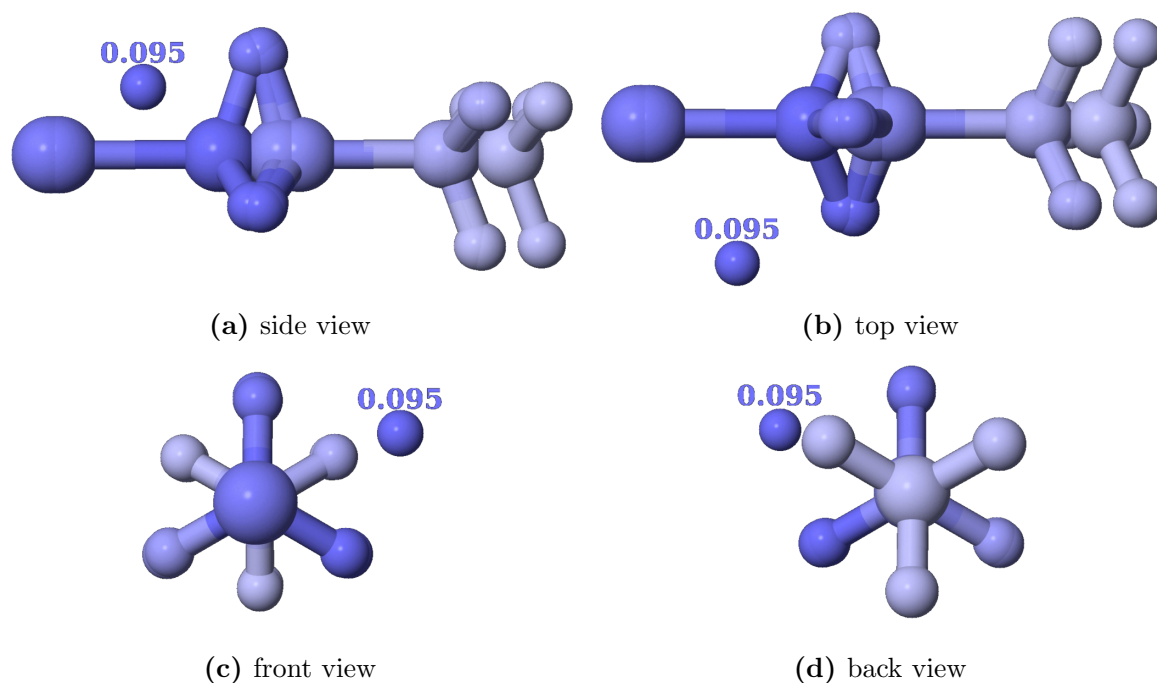


Fig. S8: Single GOCAT individual: four different views of r_{4907} (of c_2) for the $N_{Ch} = 1$ case; for illustration details, see Fig. S7.

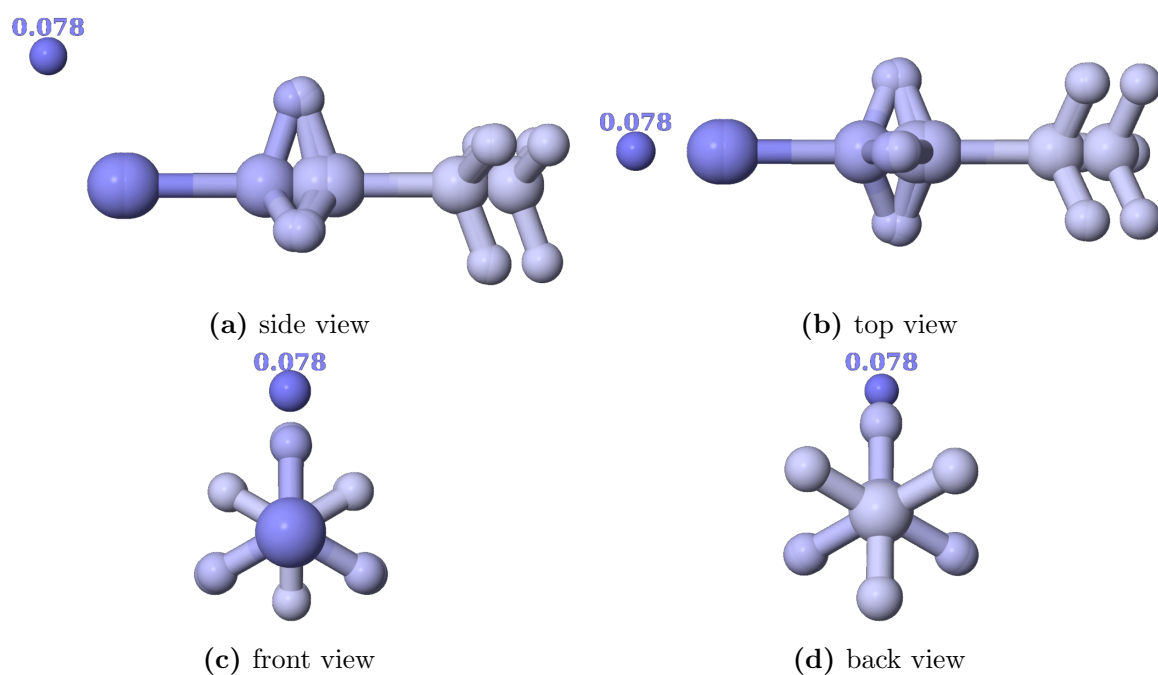


Fig. S9: Single GOCAT individual: four different views of r2471 (of c3) for the $N_{\text{Ch}} = 1$ case; for illustration details, see Fig. S7.

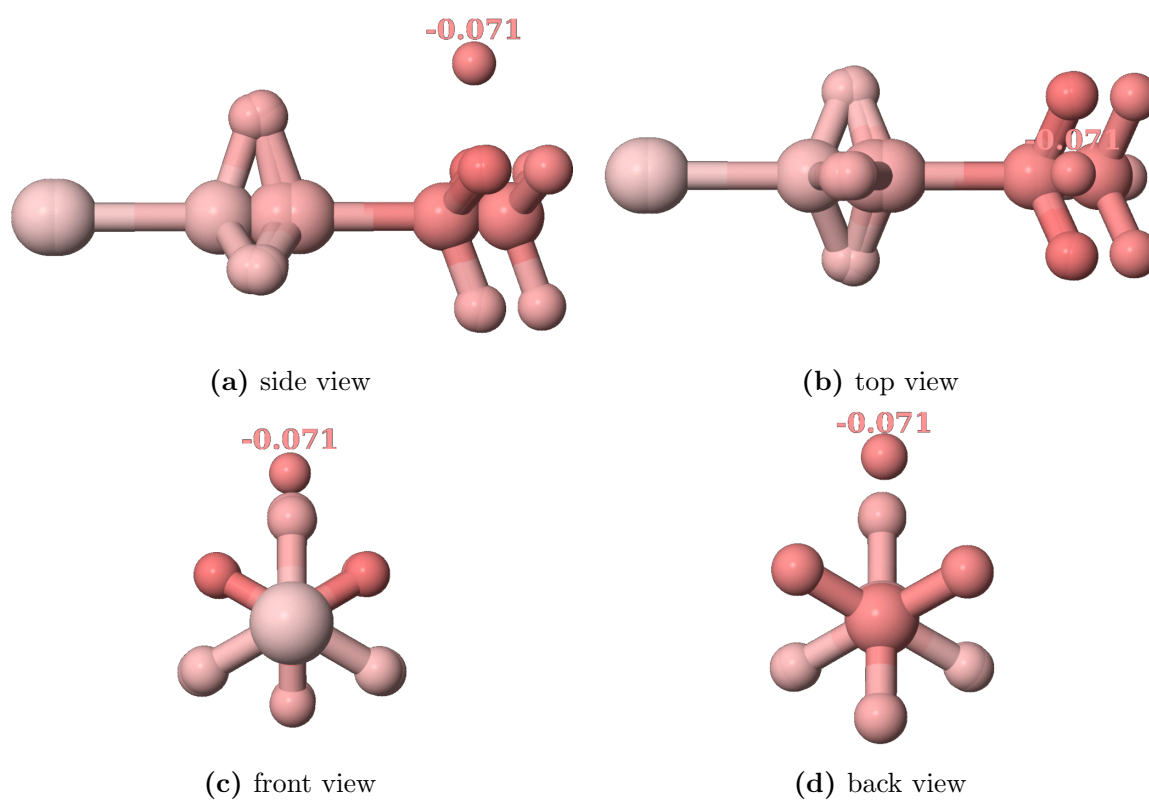


Fig. S10: Single GOCAT individual: four different views of r1694 (of c12) for the $N_{\text{Ch}} = 1$ case; for illustration details, see Fig. S7.

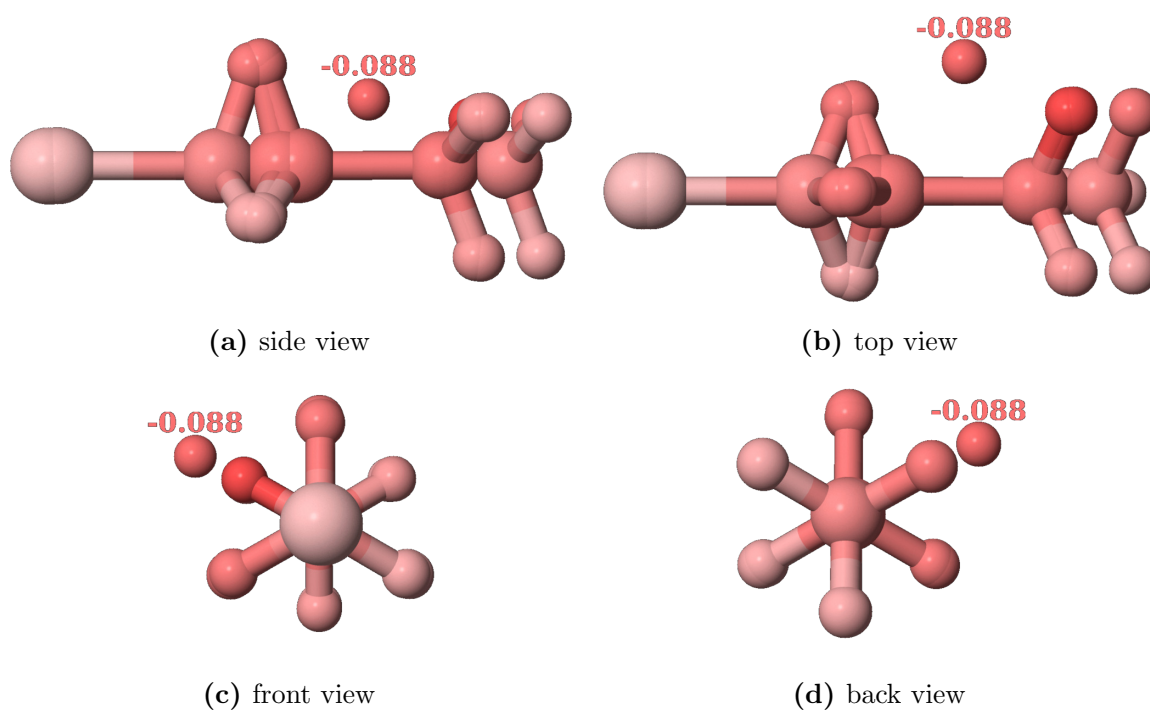


Fig. S11: Single GOCAT individual: four different views of r1366 (of c13) for the $N_{\text{Ch}} = 1$ case; for illustration details, see Fig. S7.

S3 PM7: $N_{\text{Ch}} = 3$ case (non-neutral summed charges)

S3.1 Cluster Analysis

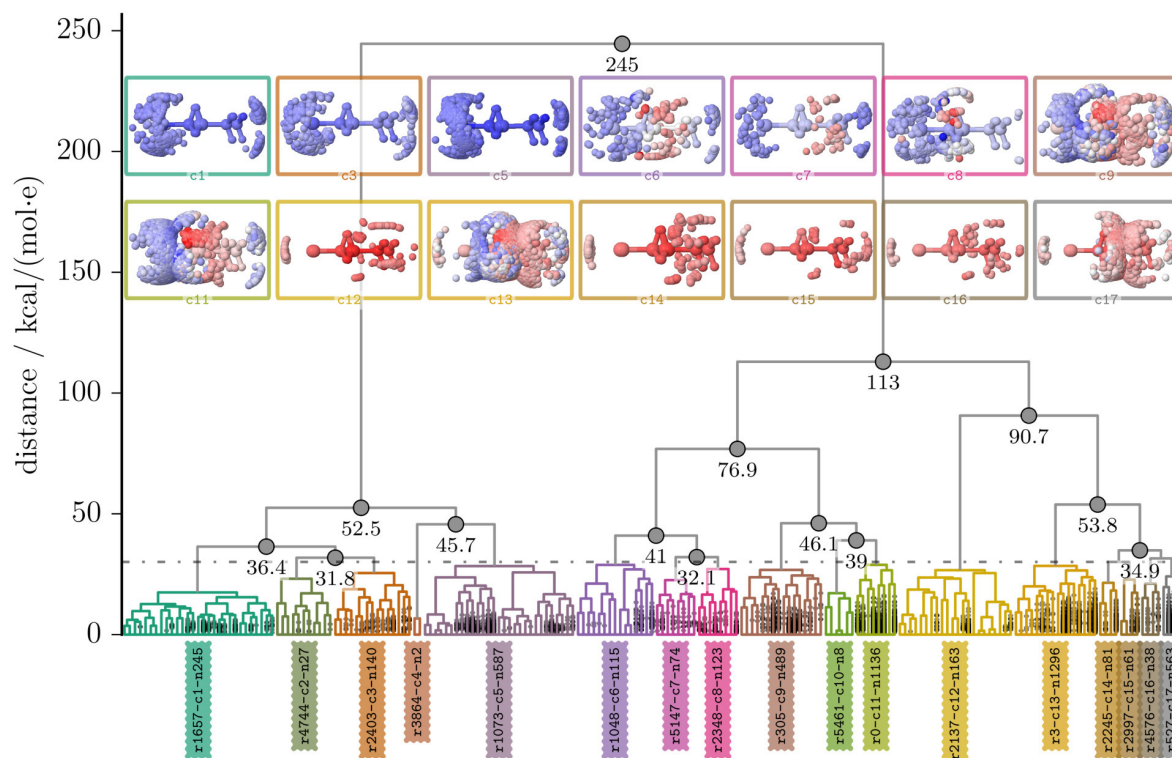


Fig. S12: $N_{\text{Ch}} = 3$ GOCATs (PM7): Dendrogram of final database with 5148 non-identical individuals using the average linkage strategy (unweighted pair group method with averaging, UPGMA). As distance metric, Eq. (2) based on Eq. (1) (main article) was used. For illustration purposes the dendrogram was cut to reach 17 different clusters, effectively cutting at a distance of (directly below) $30.09 \text{ kcal mol}^{-1} \text{e}^{-1}$ (dotted line). As outliers (very small clusters) are also filtered out this way (i.e., build separate small clusters in the dendrogram), some selected examples of bigger clusters (main funnel on fitness surface) are shown by corresponding overlain individuals of that cluster. This dendrogram is truncated, i.e., the small ellipses at the leaves show additional branching points of the binary tree that are not plotted. The main branching points of the clusters are separately annotated by the numbered dots.

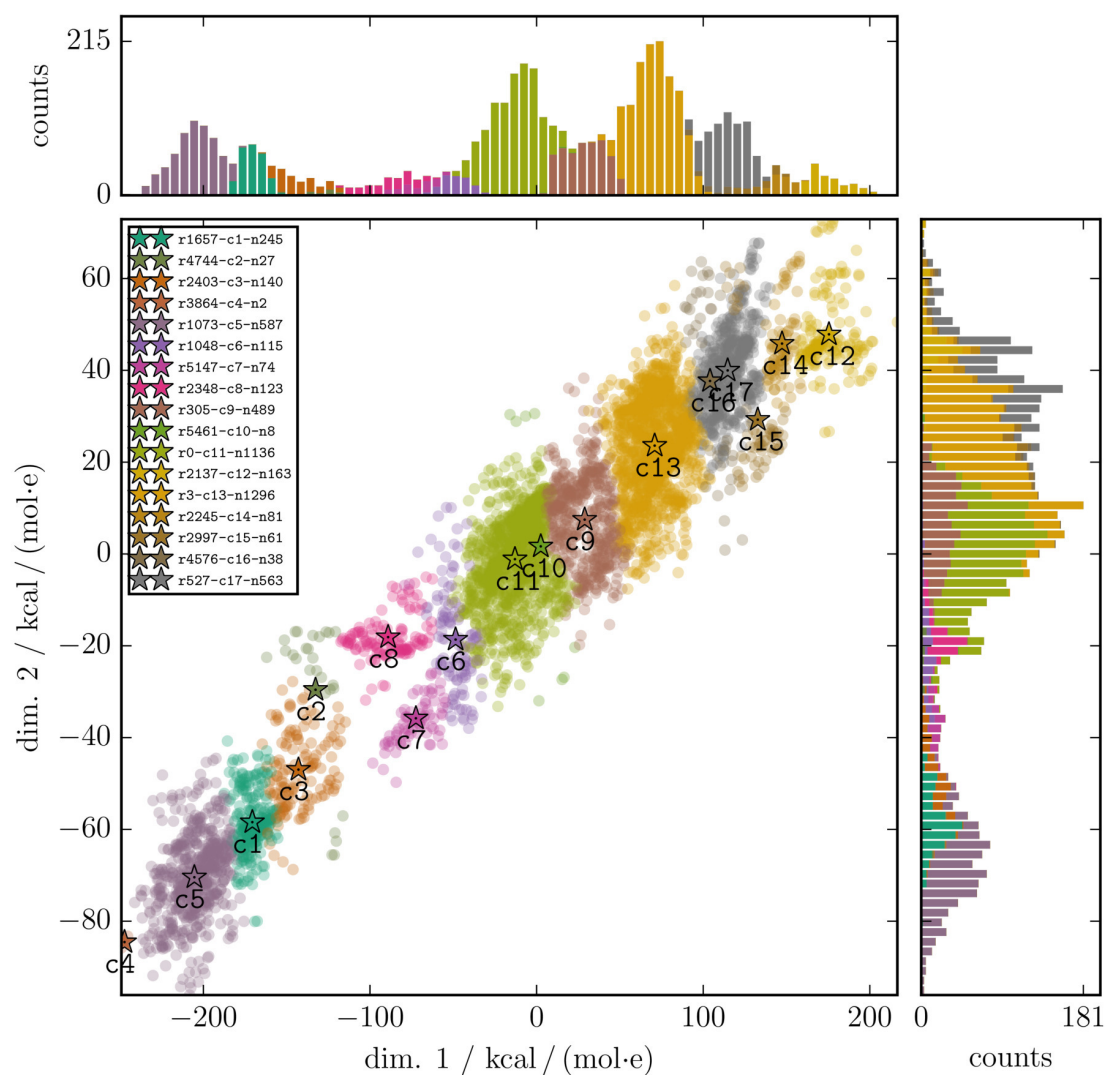


Fig. S13: $N_{Ch} = 3$ GOCATs (PM7): Multidimensional Scaling as 2D projection of the higher dimensional ESP-distance data is shown for illustrating the core cluster regions. Colored stars are the calculated mean 2D-coordinates (centroid) of a cluster. The (unnormalized) stacked histograms illustrate the number of individuals in those clusters.

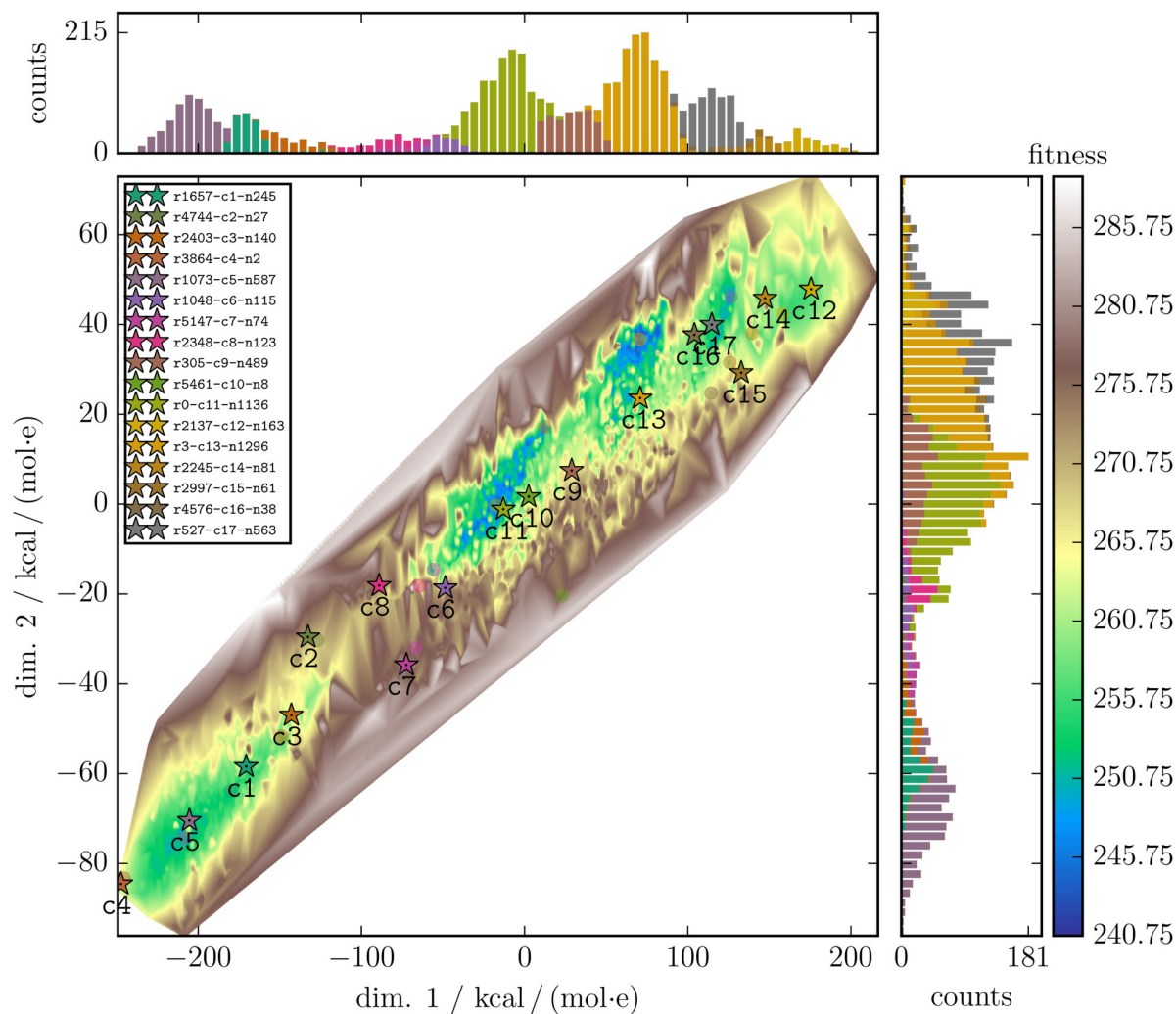


Fig. S14: $N_{\text{Ch}} = 3$ GOCATs (PM7): Multidimensional Scaling as 2D projection similar to Fig. S13; for illustration details, see Fig. S1. The final fitness surface is already quite rugged, shown by the smallest fitness values that could be found after all the independent runs and the complete local re-optimization at the end of the complete database. At least 4 different broader minima regions (around c5, c11, c13, c17, including many more subtle sub-minima) can be detected as main “basins of attraction” or funnels of the fitness surface where the biggest clusters and thus most local minima can be distributed.

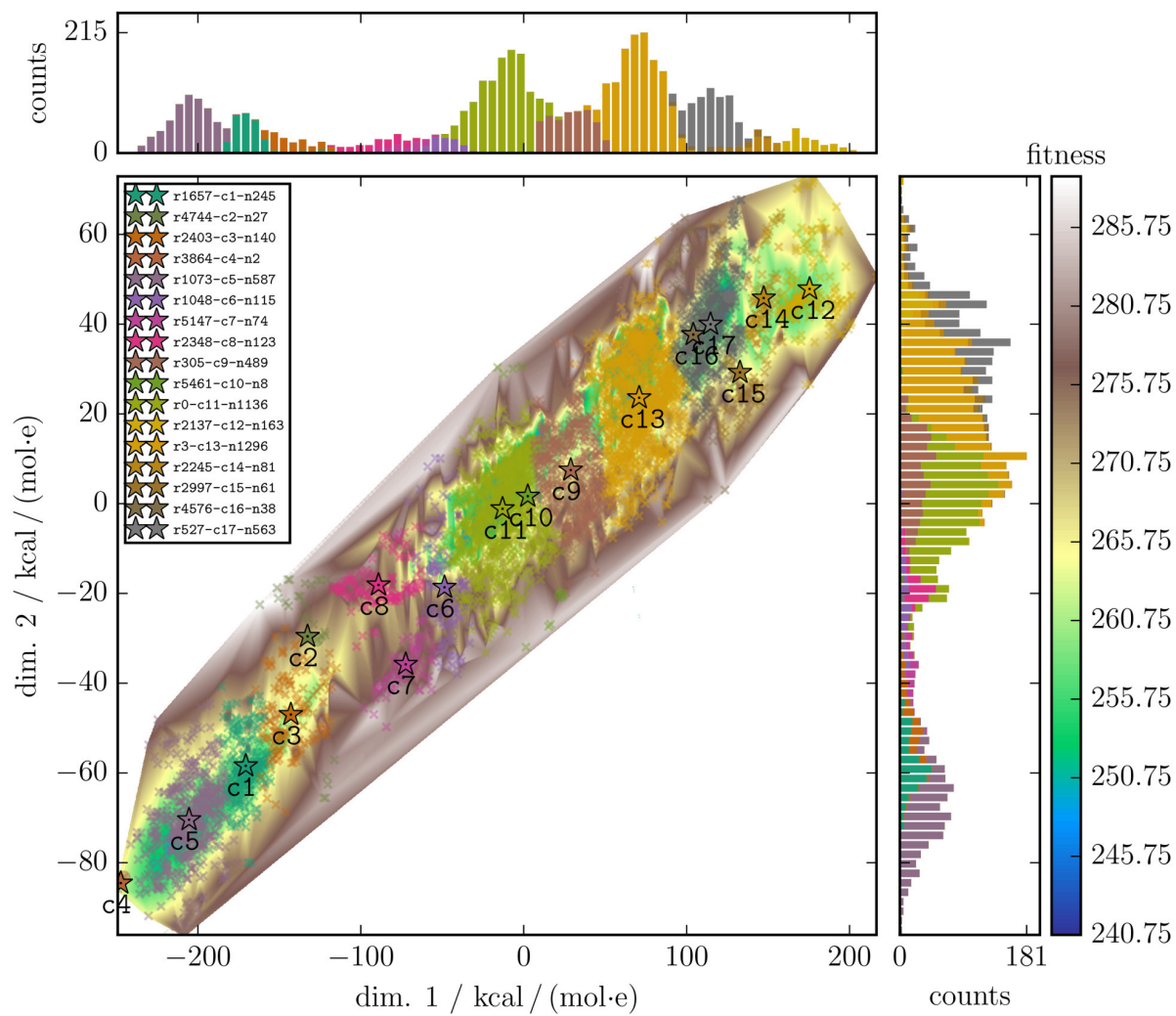


Fig. S15: $N_{Ch} = 3$ GOCATs (PM7): Similar plot to Fig. S14 including data points; for illustration details, see Fig. S2.

S3.2 Reaction Paths (Selected Clusters)

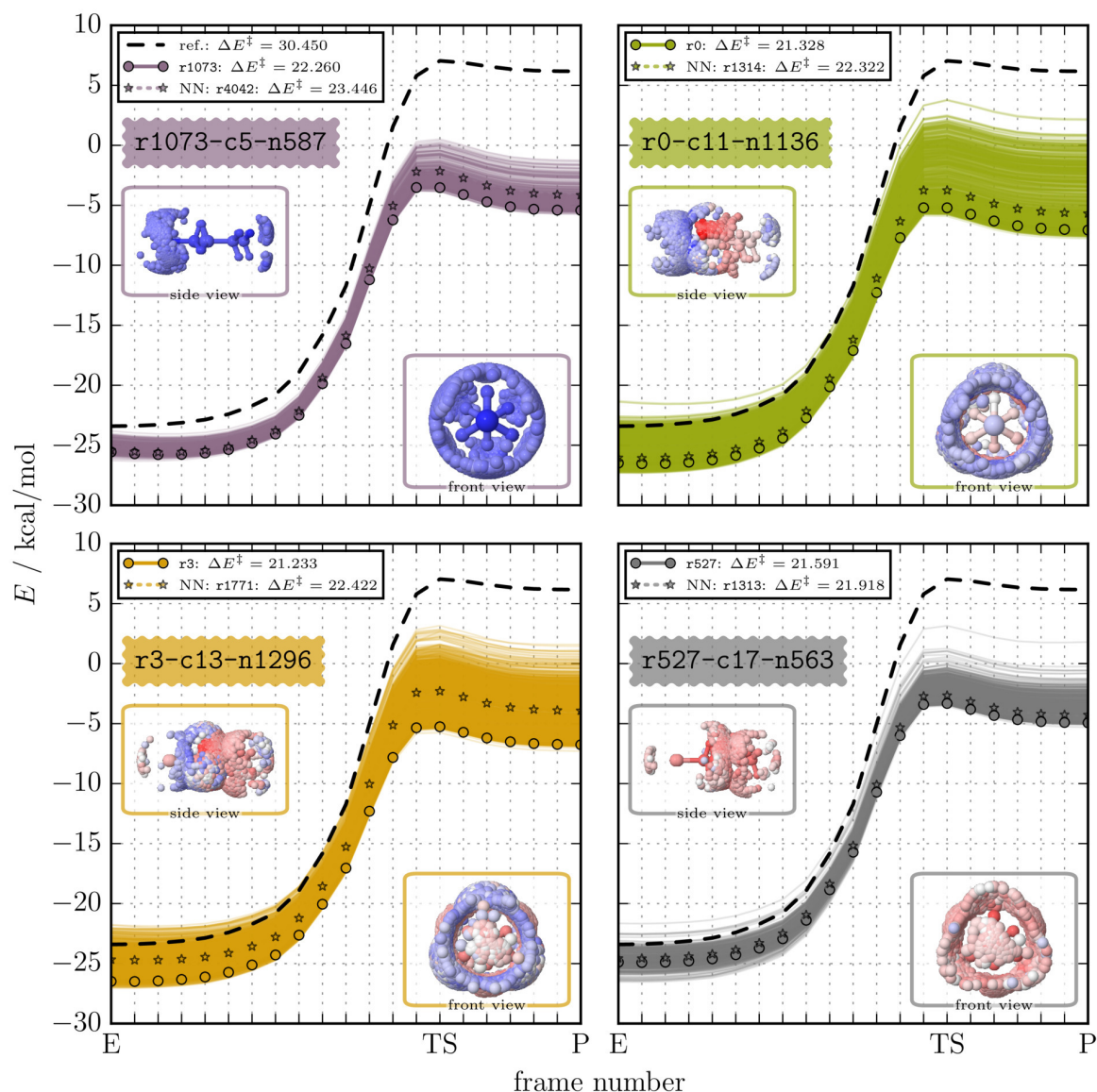


Fig. S16: $N_{\text{Ch}} = 3$ GOCATs (PM7): (Pre-optimized) reaction path for the PM7 reference calculation is shown as well as the path within the electrostatic GOCATs of the four main clusters (c5, c11, c13, c17, compare with Figs. S12 to S14); for illustration details, see Fig. S3. Color values between $[-0.268, +0.268]$ e for charges and $[-59.721, +59.721]$ kcal mol $^{-1}$ e $^{-1}$ for ESP values. Note, that the “ring-structure” of the $N_{\text{Ch}} = 1$ case discussed in the main article in that cluster $c3_{N_{\text{Ch}}=1}$ appears again in c5 here in the $N_{\text{Ch}} = 3$ case with additional positive domains at the ammonia H-atoms. c17 is similar to $c9_{N_{\text{Ch}}=1}$ as negative cluster, and the almost neutral c11 includes both these corner cases. When going to the $N_{\text{Ch}} = 10$ case, most clusters there will be similar to this $c11_{N_{\text{Ch}}=3}$ here, see Fig. S24: This is the main trend/convergence of the GOCATs with increasing number of charges discussed in the main article.

S3.3 Selected Details

Table S2: Properties of the final clusters for the $N_{\text{Ch}} = 3$ case and of their best rank individual: For clusters, both mean values of those clusters as well as standard deviations (parentheses) are given. For separate individuals, just their single value is presented.

cluster name or individual	fitness	barrier [kcal mol ⁻¹]	grad. norm (E) [kcal mol ⁻¹ Å ⁻¹]	grad. norm (TS) [kcal mol ⁻¹ Å ⁻¹]	grad. norm (P) [kcal mol ⁻¹ Å ⁻¹]	sum. charge (elem. charge)
r1657-c1-n245	259.114(5.479)	23.096(0.477)	8.435(0.771)	11.631(0.185)	6.763(0.979)	0.276(0.013)
r4744-c2-n27	279.594(5.715)	24.939(0.445)	4.669(0.408)	11.666(0.212)	7.551(0.737)	0.177(0.016)
r2403-c3-n140	265.249(6.108)	23.686(0.533)	7.073(1.428)	11.560(0.192)	7.187(1.359)	0.223(0.022)
r3864-c4-n2	262.577(0.574)	22.492(0.013)	12.086(0.108)	11.940(0.056)	7.655(0.322)	0.409(0.000)
r1073-c5-n587	256.645(6.272)	22.777(0.550)	10.107(0.550)	11.745(0.213)	7.089(0.998)	0.337(0.022)
r1048-c6-n115	270.948(8.406)	24.136(0.830)	7.330(1.448)	11.630(0.259)	9.504(1.053)	0.064(0.016)
r5147-c7-n74	276.880(3.064)	24.667(0.354)	8.909(0.892)	11.599(0.319)	10.355(0.457)	0.122(0.020)
r2348-c8-n123	272.421(5.449)	24.346(0.513)	4.629(1.134)	11.548(0.198)	7.214(1.301)	0.111(0.021)
r305-c9-n489	263.602(8.242)	23.443(0.801)	7.998(1.189)	11.695(0.239)	9.658(0.870)	-0.063(0.021)
r5461-c10-n8	282.052(3.720)	25.005(0.377)	7.100(0.714)	11.914(0.156)	9.315(0.484)	-0.043(0.012)
r0-c11-n1136	255.421(11.008)	22.665(1.062)	8.372(1.727)	11.765(0.215)	9.359(0.846)	0.004(0.023)
r2137-c12-n163	260.603(5.875)	22.920(0.449)	10.072(0.534)	11.859(0.272)	10.694(0.534)	-0.302(0.022)
r3-c13-n1296	256.963(9.369)	22.763(0.901)	8.641(1.093)	11.827(0.236)	9.805(0.854)	-0.129(0.022)
r2245-c14-n81	260.591(4.101)	22.990(0.308)	8.930(0.534)	11.836(0.275)	10.582(0.578)	-0.254(0.009)
r2997-c15-n61	268.073(4.116)	23.872(0.406)	9.937(0.498)	11.670(0.201)	9.430(0.718)	-0.244(0.013)
r4576-c16-n38	275.643(4.987)	24.467(0.437)	8.953(1.058)	11.831(0.243)	8.781(0.896)	-0.208(0.014)
r527-c17-n563	256.705(6.402)	22.714(0.650)	9.260(0.813)	11.876(0.238)	9.941(0.502)	-0.198(0.018)
r1657	252.438	22.706	9.221	11.195	5.680	0.299
r4744	267.808	24.007	4.201	11.413	6.181	0.173
r2403	255.875	23.014	6.677	11.207	5.189	0.236
r3864	262.171	22.483	12.009	11.980	7.428	0.410
r1073	249.602	22.260	10.260	11.572	5.959	0.348
r1048	249.480	21.669	10.201	12.433	10.028	0.070
r5147	271.545	24.293	8.828	11.522	9.998	0.112
r2348	255.685	22.951	6.415	11.314	5.671	0.080
r305	244.283	21.522	10.184	11.973	10.334	-0.038
r5461	274.621	24.203	6.888	12.093	10.329	-0.050
r0	240.794	21.328	10.473	11.757	9.530	0.005
r2137	254.877	22.680	9.833	11.664	10.350	-0.270
r3	240.909	21.233	9.951	11.926	10.473	-0.126
r2245	255.256	22.606	9.277	11.876	10.093	-0.238
r2997	257.903	22.744	9.802	12.031	10.082	-0.216
r4576	266.701	23.875	8.469	11.477	9.920	-0.206
r527	245.756	21.591	10.082	12.096	10.144	-0.213

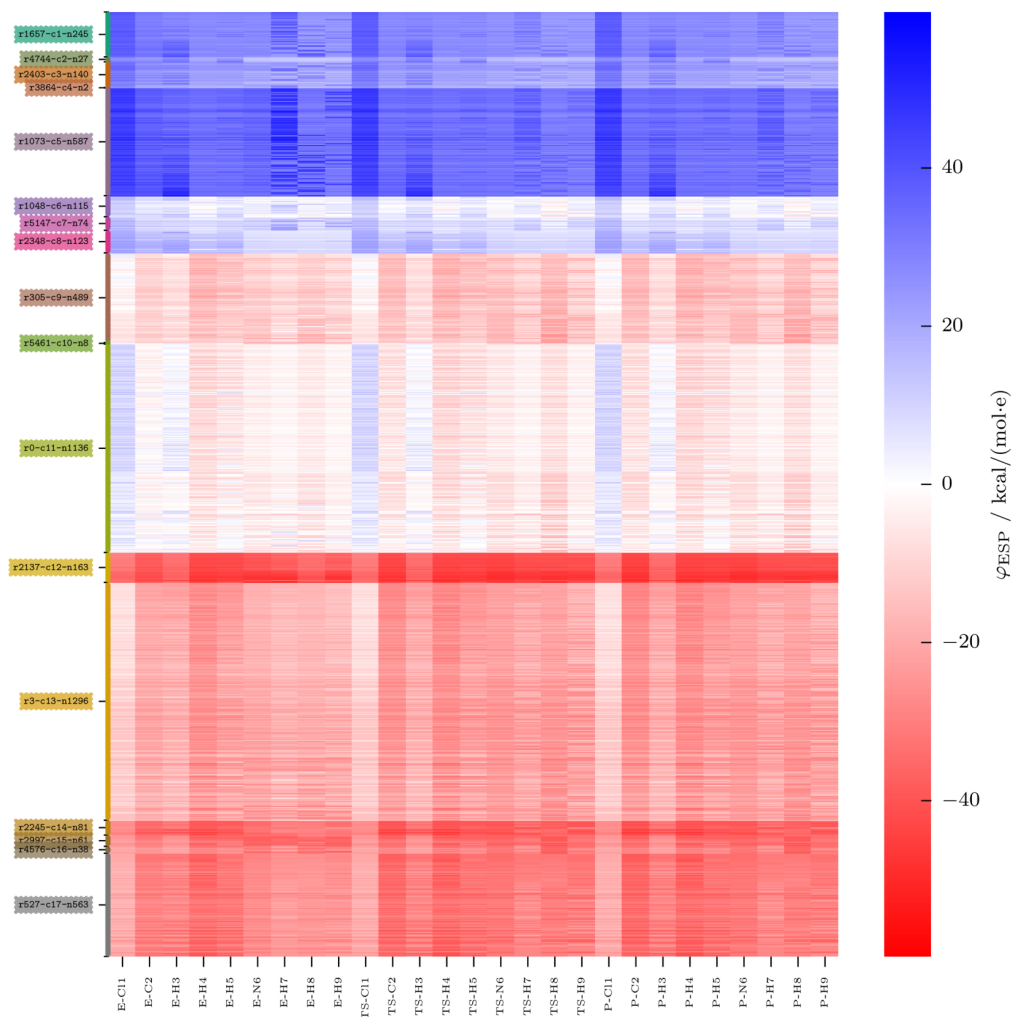


Fig. S17: $N_{\text{Ch}} = 3$ GOCATs (PM7): Heatmap of the complete database chunked into 17 clusters; for illustration details, see Fig. S6.

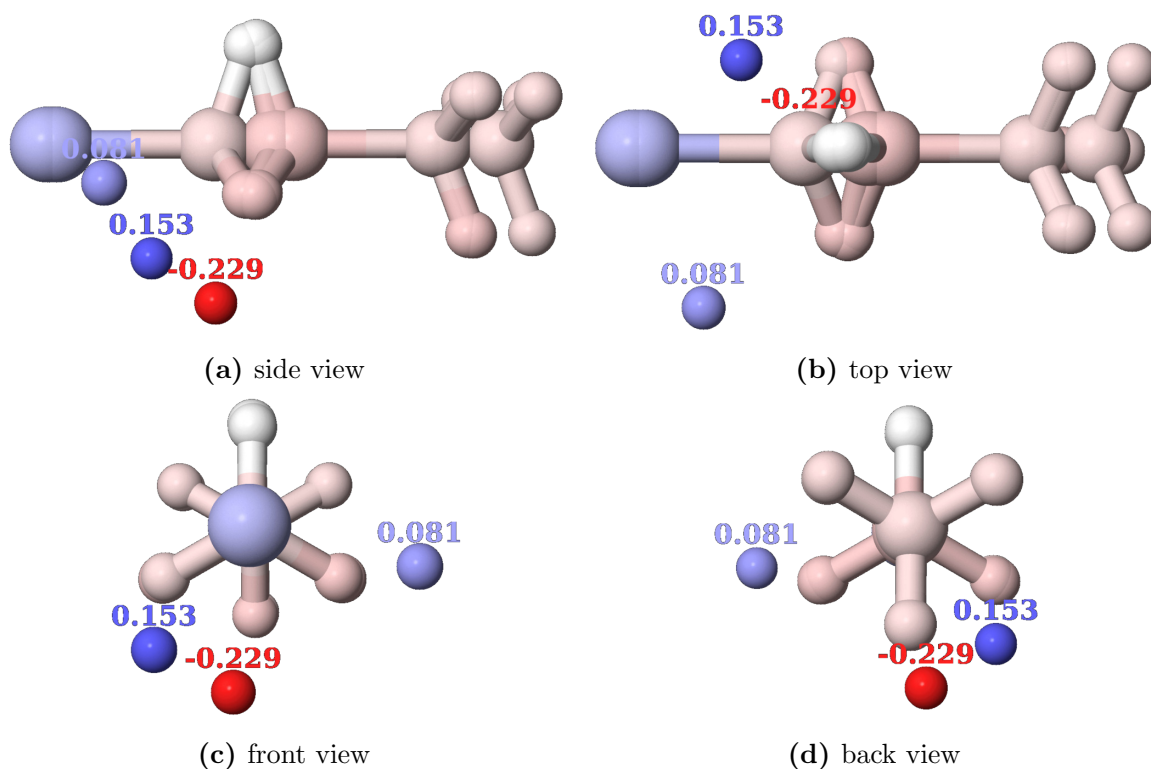


Fig. S18: Single GOCAT individual: four different views of r_0 (of c_{11}) for the $N_{\text{Ch}} = 3$ case with values given for the partial charges. Both partial charges and the atoms of the core frames (E, TS, P shown) are colored red/blue in the ranges $[-0.268, +0.268] e$ for charges and $[-59.721, +59.721] \text{ kcal mol}^{-1} e^{-1}$ for ESP values.

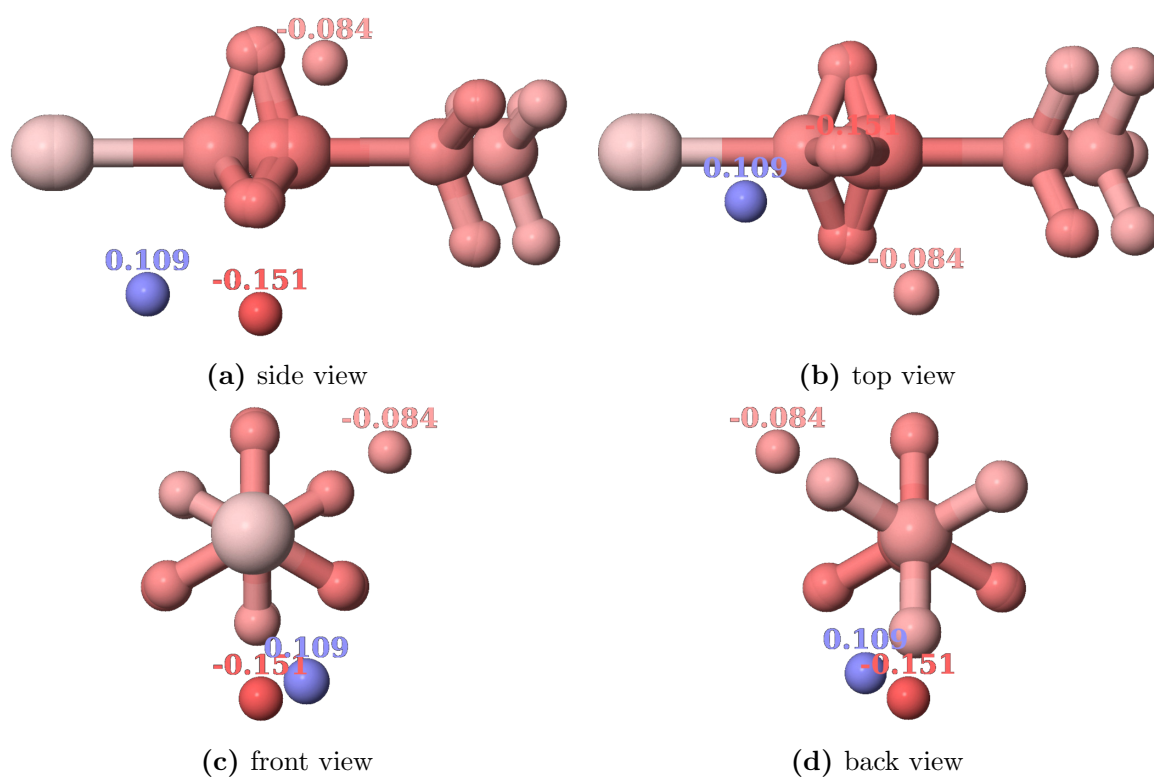


Fig. S19: Single GOCAT individual: four different views of r_3 (of c_{13}) for the $N_{Ch} = 3$ case; for illustration details, see Fig. S18.

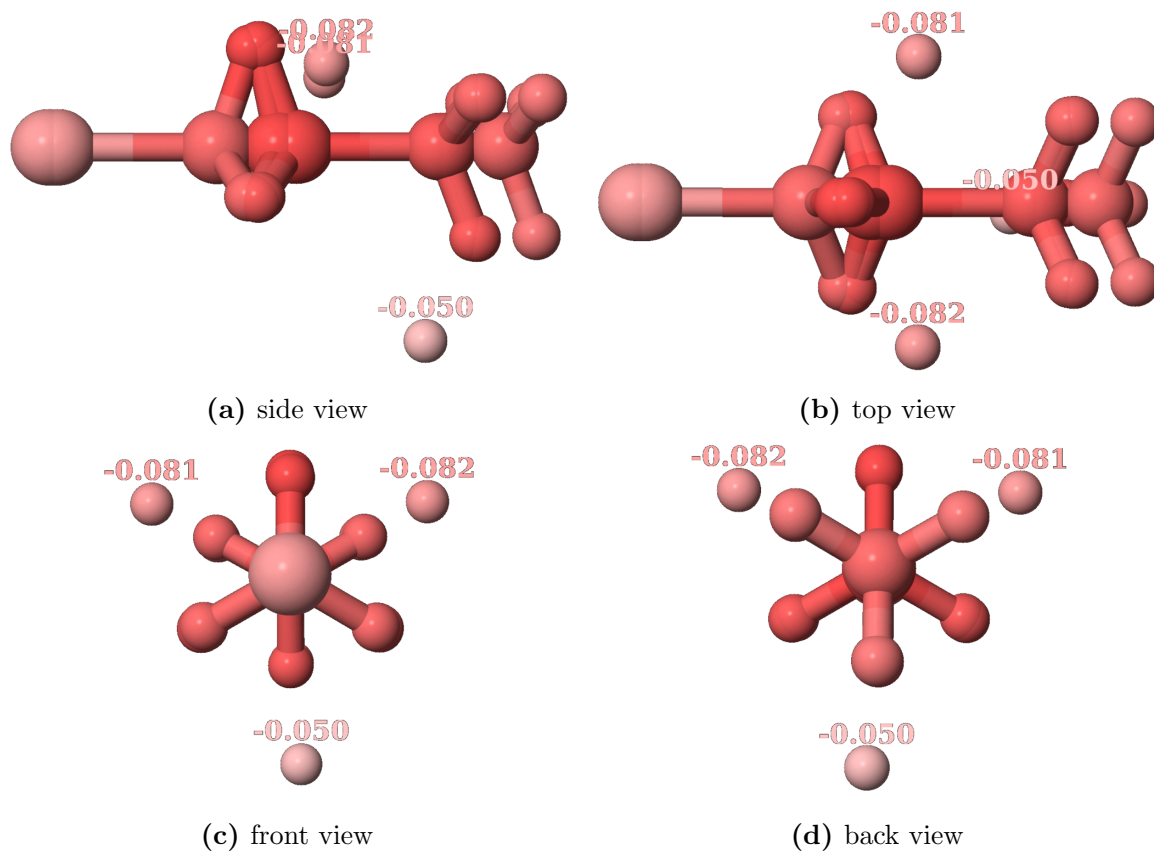


Fig. S20: Single GOCAT individual: four different views of r527 (of c17) for the $N_{\text{Ch}} = 3$ case; for illustration details, see Fig. S18.

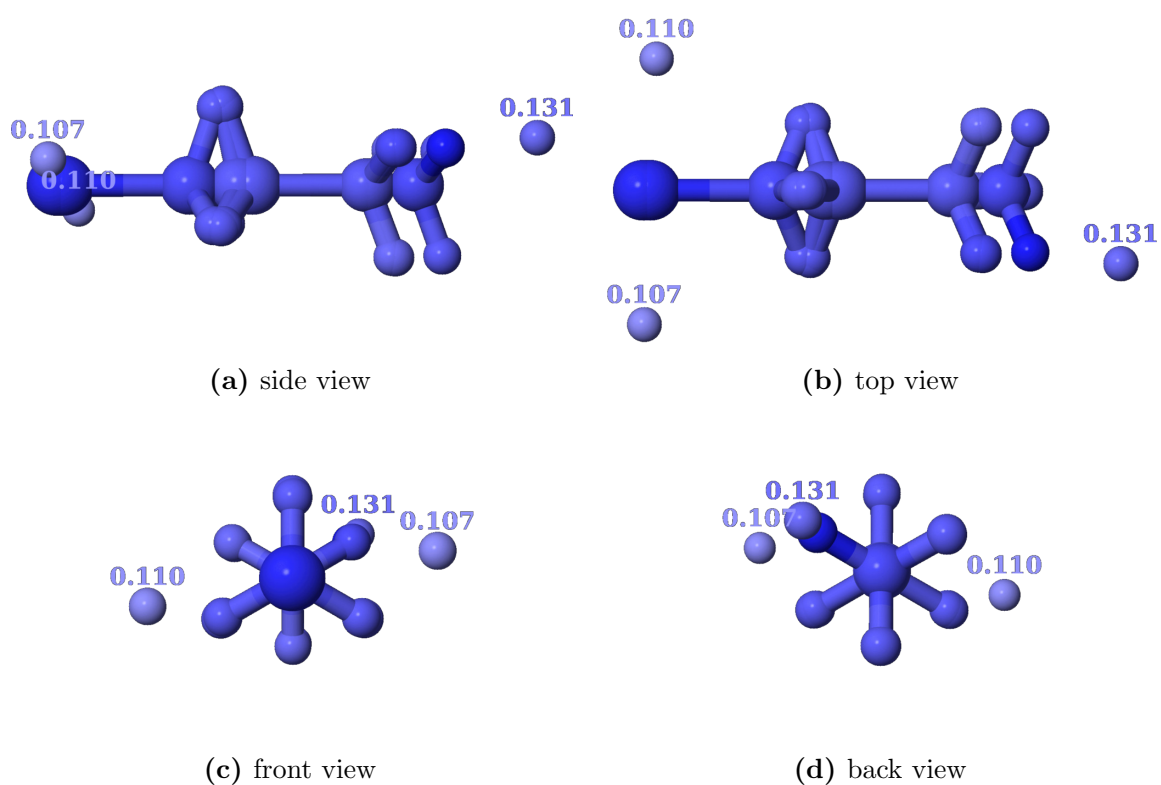


Fig. S21: Single GOCAT individual: for different views of r1073 (of c5) for the $N_{\text{Ch}} = 3$ case; for illustration details, see Fig. S18.

S4 PM7: $N_{\text{Ch}} = 10$ case (non-neutral summed charges)

S4.1 Complementary Discussions

Going to multiple partial charges as abstract GOCATs increases the complexity of the optimization but also allows to create electrostatic embeddings that address the core-atoms separately, providing some or all of them with individually optimized surroundings, and that may also produce correlated or synergistic effects. We have tried several different N_{Ch} values but present here only the $N_{\text{Ch}} = 10$ size. A MDS plot for this case is given in Fig. S22. Note that the final summed charge of all 10 partial charges is not constrained here (in contrast to the next Section), such that highly charged GOCATs are (yet) allowed, presuming that another shell of a real molecular embedding might compensate for that in a concrete system. Additional observations to the ones mentioned above are (compare also with complementary figures below for this GOCAT size):

- More non-redundant GOCATs are possible (bigger search space due to higher complexity/more freedom, and accordingly the already bigger database shown will not be an exhaustive representation of the bigger search space in this case).
- GOCATs happen to lie on a straight line, i.e., the database might even be reduced to *one* (essential) dimension. This provides a hint that one specific qualitative solution (ESP relation on the atoms) is most important for the catalytic effect and that now embeddings with enough freedom or flexibility (of 10 charges) can discover this.
- Very broad ESP range, although the biggest clusters build up at the center, which also maps to mainly overall neutral GOCATs and corresponds to broad basins of attraction (funnels) on the fitness surface.

This trend starts already in smaller GOCAT cases (compare with the $N_{\text{Ch}} = 3$ case, illustrated in Section S3 as an “intermediate” case with respect to complexity). Additionally, in most clusters some highly specific (needle-like) minima can be found, which are also

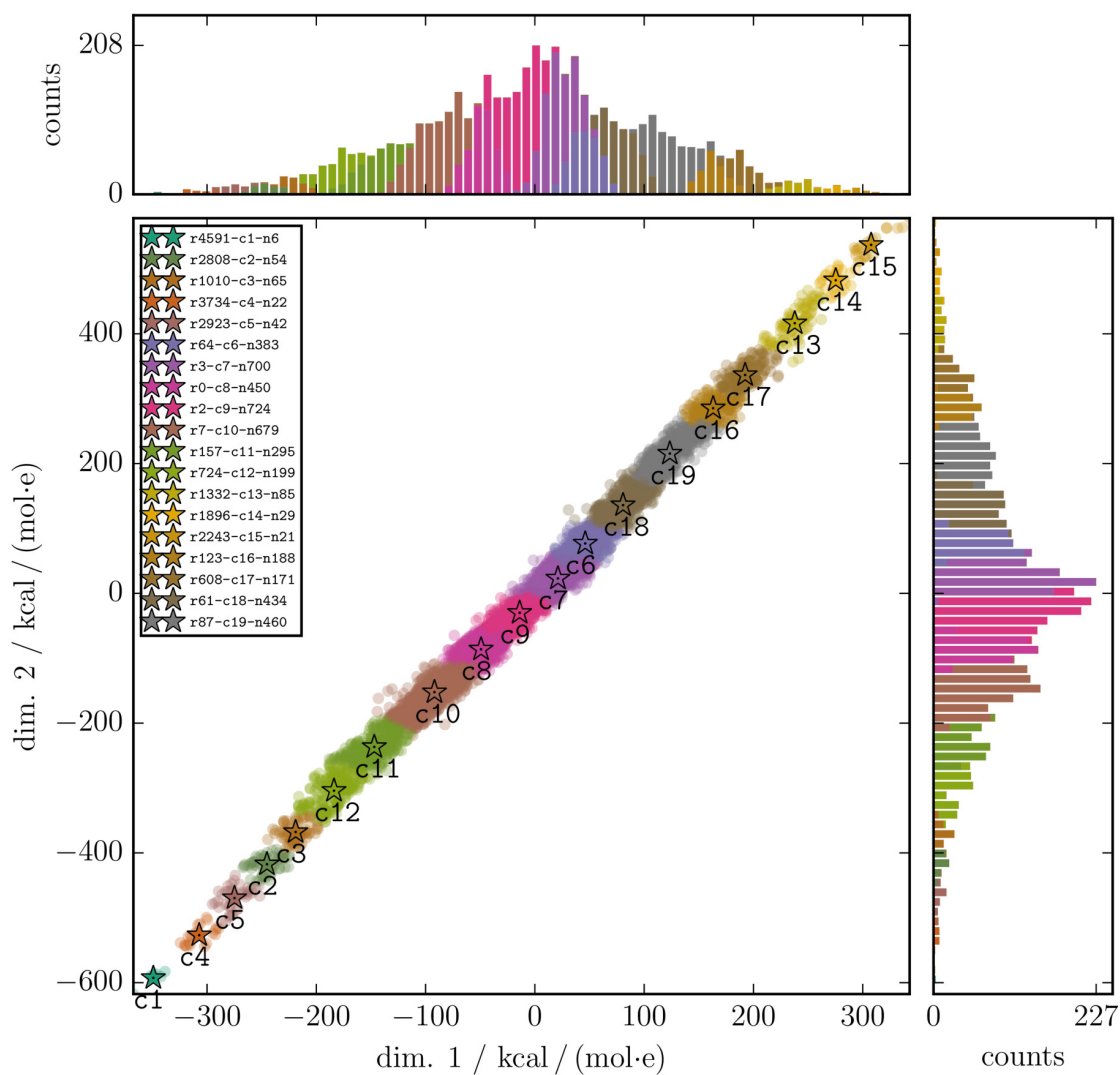


Fig. S22: $N_{\text{Ch}} = 10$ GOCATs (PM7, non-neutral): MDS 2D projection of the higher dimensional ESP-distance of a database of 5007 non-identical GOCATs; for illustration details, compare with Fig. S2 or Fig. S1.

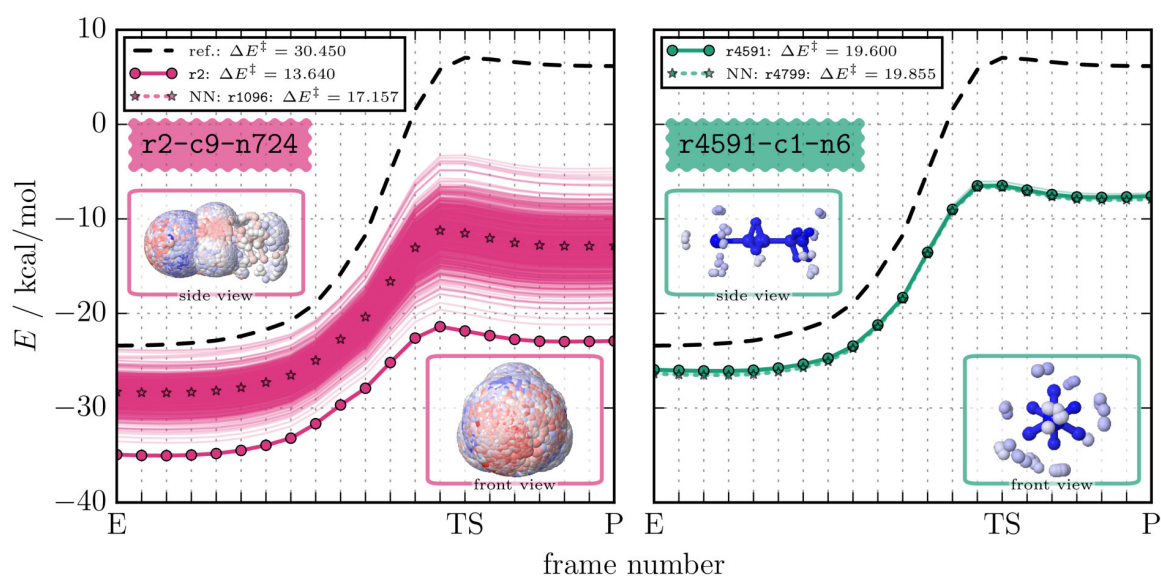


Fig. S23: $N_{\text{Ch}} = 10$ GOCATs (PM7, non-neutral): Reaction paths of one big (almost neutral embedding) GOCAT cluster (c9) and an outlier one (c1) are shown. For illustration details, see Fig. S3. Note, that the overlain GOCATs in the insets of the plots are misleading, as almost a continuous embedding is shown as superposition of, e.g., each $724 \cdot 10$ charges in the c9 case.

apparent in the corresponding reaction profiles that are given in Fig. S23. There, the best individual (r2) of that cluster c9 shown is quite unique and lies also energetically a bit offside (line with dotted points). The reason for this is illustrated in the main article, and is caused mainly by very symmetric or “perfectly” tuned Cartesian placements of partial charges to reach the essential ESP values. Of course, it would always be possible to increase the number of clusters during hierarchical clustering in order to resolve those details and also to look at minor qualitative differences within on single cluster. c9 is chosen for illustration because it is almost neutral and the biggest cluster in this setting with a mean summed charge of all the GOCAT charges of $-0.034e$, whereas c1 illustrates a very positive outlier cluster that is mapped to the lower left corner region in the MDS plot. This encodes one extreme case of the electrostatic potential GOCATs found and has a mean summed charge of about $+1.038e$ (the other extreme would be c15 with $-1.108e$). But in each case, the final reaction profile’s characteristics already observed in the $N_{\text{Ch}} = 1$ case above are amplified here. The overall absolute stabilization energies increase, leading to even lower barriers of

e.g. $\Delta E^\ddagger = 13.6 \text{ kcal mol}^{-1}$ for **r2**. Also the highest-energy frame (within the discretization of pre-calculated frames) tries to shift to the educt, which is restrained partially. This could be interpreted in concordance with Hammond's postulate^{S5} that there is a late TS in the gas-phase that will shift to the educt with increased stabilization due to the surrounding that might at the end also reach exothermicity. And from this same influence to an effective PES also the catalytic effect may follow, i.e. smaller reaction barriers.

Note that the shown case of $N_{\text{Ch}} = 10$ is already a converged end point of a general trend. For example, also bigger GOCATs like $N_{\text{Ch}} = 20$ (not shown) lead to overall very similar embeddings in terms of ESP values (apart from bigger noise introduced by the higher complexity of the optimization problem in that case). Already with the $N_{\text{Ch}} = 3$ case (given above) the linearization of the database, the huge ESP range corresponding to highly negatively or positively charged GOCATs and the trend in the reaction profiles start. Moreover, also there is of course no bijectivity of Cartesian placement of partial charges and the resulting ESP at the core atoms anymore (which is also the reason for the additional BoB descriptor used in the filtering process). This results in different possible Cartesian domains on the vdW surface for similar GOCATs: Sometimes erratically different charges (e.g. see the Cl-atom site of **r2-c9-n724** in Fig. S23 which carries both some very positive charges of one GOCAT and very negative charges of another GOCAT) can be placed at different Cartesian domains on the vdW surface in order to reach a quite similar ESP on the core atoms in a surjective manner. Therefore, they ended up in the same cluster.

Additionally, we performed PCA of that same database to describe the most prominent basis vectors of the ESP with respect to the variance of the GOCATs. Thus, up to 99.6% of the overall variance of all ESPs at the core frames is already explained by just the first PCA component. This basis vector is essentially (and trivially) a normalized unit vector. So, the main (and rough) clustering in this case first of all divides the clusters based on their total separately summed (non-neutral) charge, which is the one essential dimension seen in the MDS plots (compare the striking similarity of all overlain GOCATs – especially with regard to the Cartesian domains – in the dendrogram plot in Fig. S24).

Within the complexity of $N_{\text{Ch}} = 10$ GOCATs, the free parameter of a constant shift of the conservative Coulomb scalar potential already dominates, such that we observe mainly the same ESP *relations* and trends within each cluster. Thus, although the overall GOCAT might be very positive, as in **c1**, the trend $\varphi_{\text{ESP,C1}} > \varphi_{\text{ESP,N}} > \varphi_{\text{ESP,C}}$ will generally still hold, especially at the product (contact ion pair) site, even if everything is embedded actually negatively. To erase this artificial freedom, in the remainder of the article and this ESI an additional constraint on overall charge neutrality during GOCAT optimization is added in, i.e., we are essentially zooming into the region of **c9** or **c8** of this database within another GA and local optimization.

S4.2 Cluster Analysis

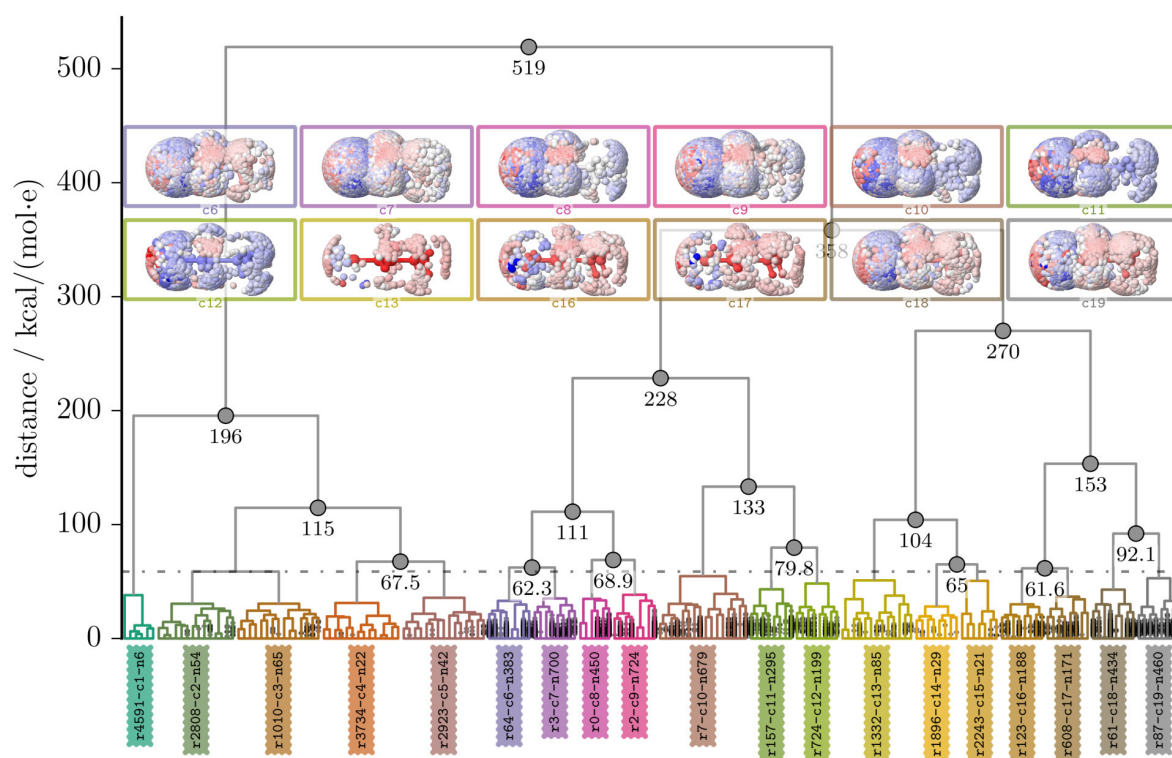


Fig. S24: $N_{\text{Ch}} = 10$ GOCATs (PM7): Dendrogram of final database with 5007 non-identical individuals using the average linkage strategy, cut into 19 different clusters, (below) $58.76 \text{ kcal mol}^{-1} \text{e}^{-1}$ (dotted line); for illustration details, see Fig. S12.

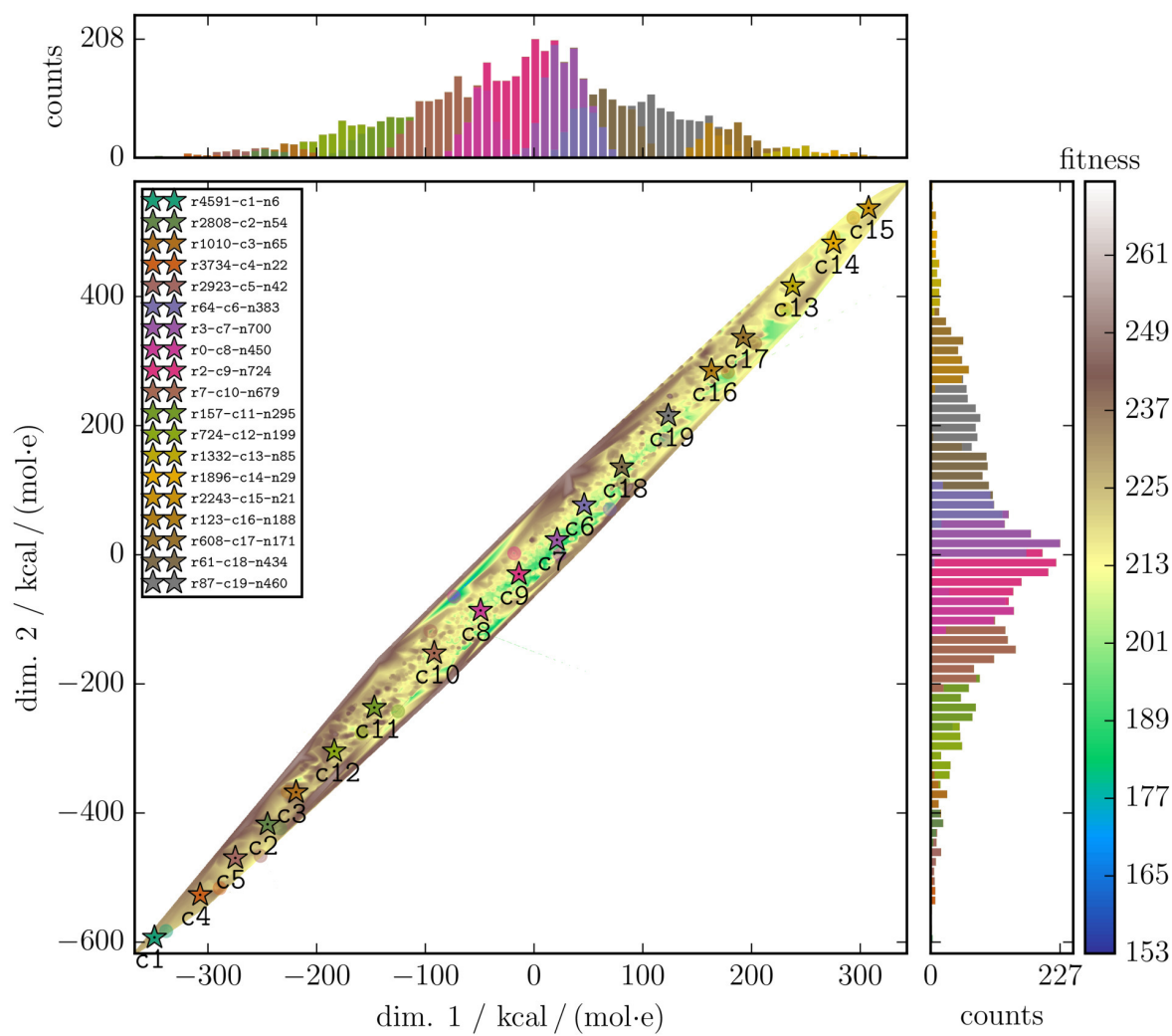


Fig. S25: $N_{\text{Ch}} = 10$ GOCATs (PM7): Multidimensional Scaling as 2D projection; for illustration details, see Fig. S1.

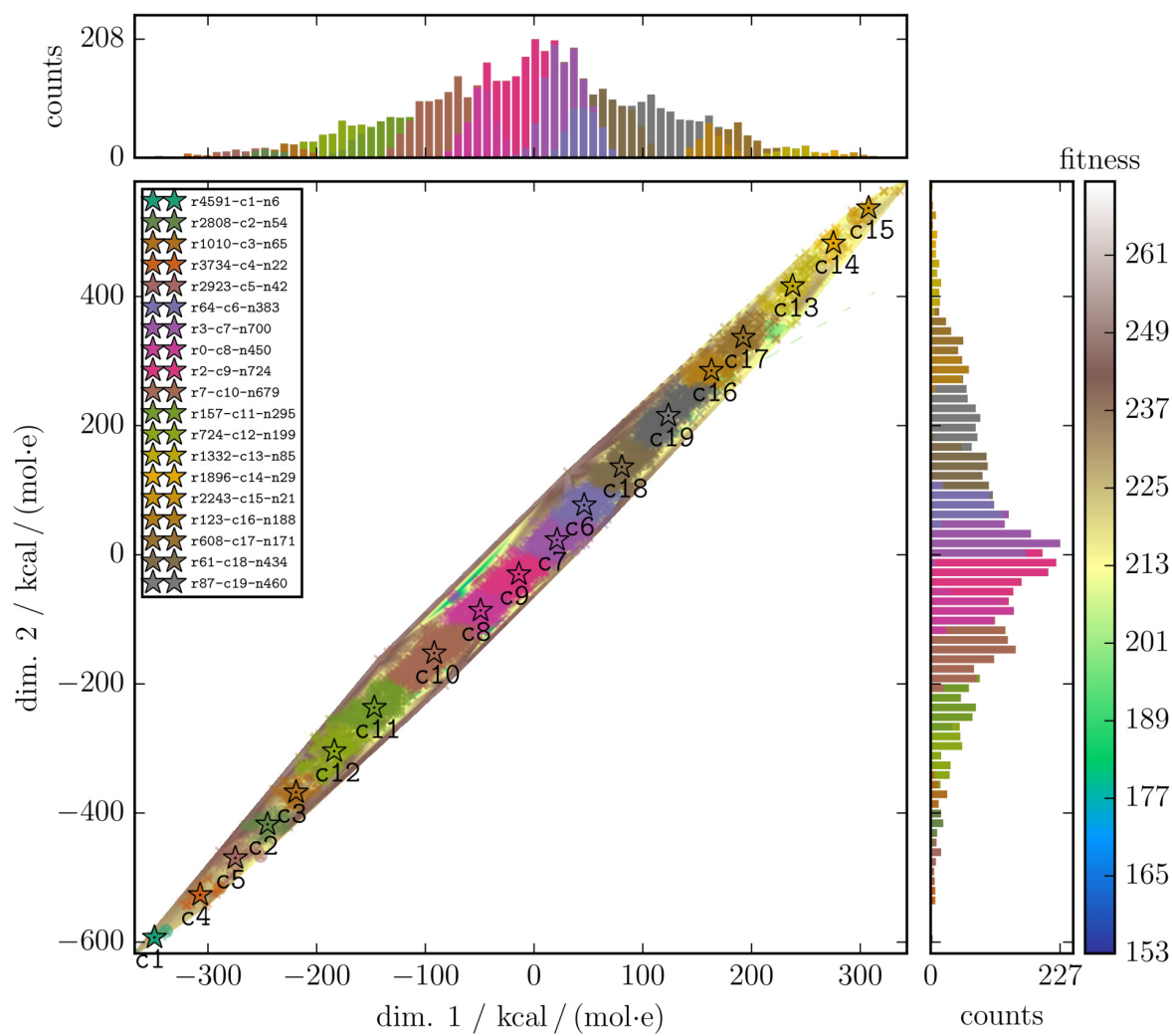


Fig. S26: $N_{\text{Ch}} = 10$ GOCATs (PM7): Similar plot to Fig. S25; for illustration details, see Fig. S2.

S4.3 Reaction Paths (Selected Clusters)

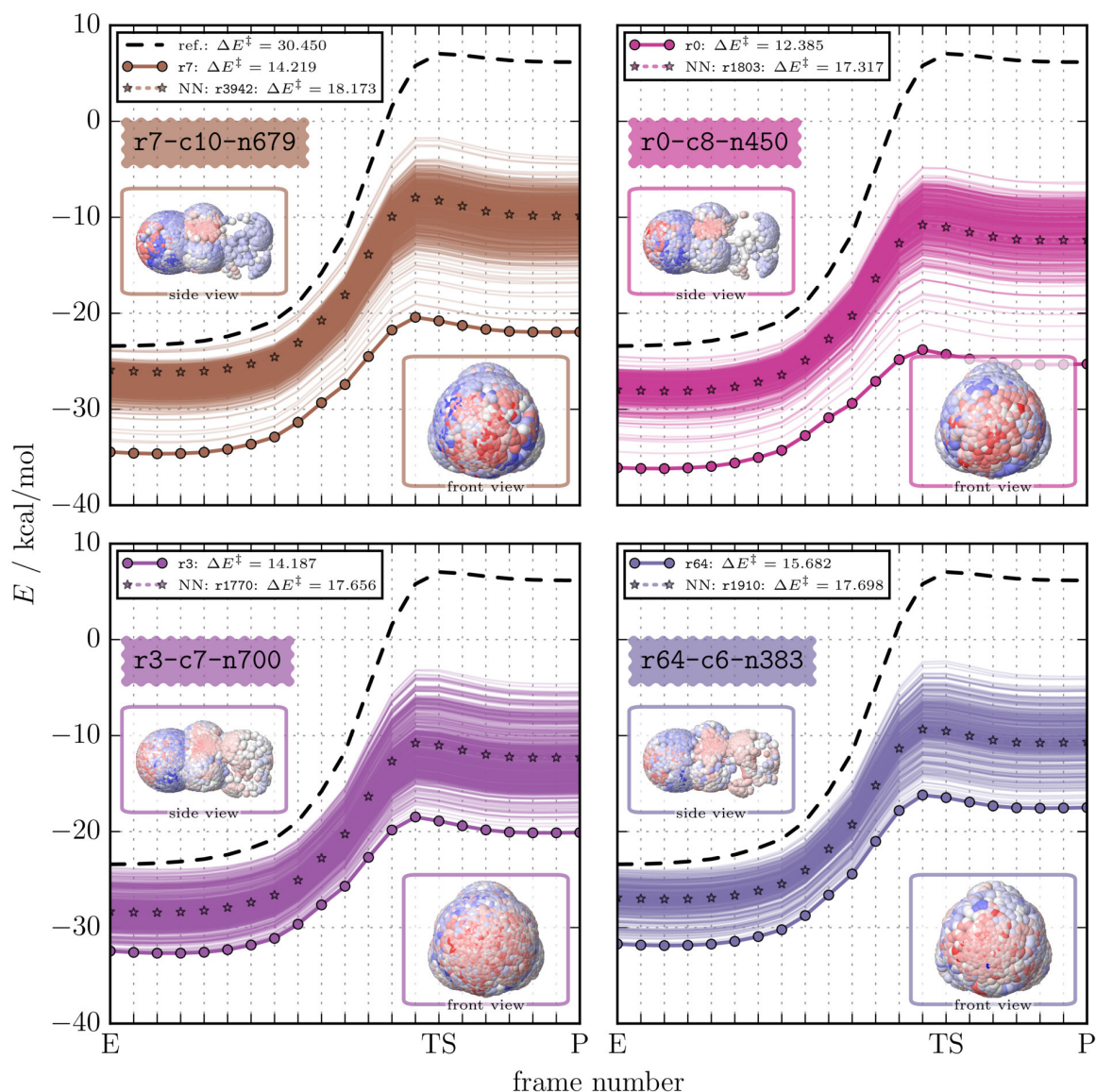


Fig. S27: $N_{\text{Ch}} = 10$ GOCATs (PM7): (Pre-optimized) reaction path for the PM7 reference calculation is shown as well as the path within the electrostatic GOCATs of four biggest clusters (c10, c8, c7, c6), from more positive ones to negative ones, see the summed charges of Table S3; for illustration details, see Fig. S3. In each case, also highly specific best ranks (lowest energy) outliers can be found and also the almost continuous Cartesian domains can be seen. Just the summed charge and thus the dominance of blue (positive) and red (negative) charges varies systematically.

S4.4 Selected Details

Table S3: Properties of the final clusters for the $N_{\text{Ch}} = 10$ case and of their best rank individual: For clusters, both mean values of those clusters as well as standard deviations (parentheses) are given. For separate individuals, just their single value is presented.

cluster name or individual	fitness	barrier [kcal mol ⁻¹]	grad. norm (E) [kcal mol ⁻¹ Å ⁻¹]	grad. norm (TS) [kcal mol ⁻¹ Å ⁻¹]	grad. norm (P) [kcal mol ⁻¹ Å ⁻¹]	sum. charge (elem. charge)
r4591-c1-n6	225.389(2.752)	19.895(0.317)	10.628(0.126)	11.753(0.171)	8.129(0.370)	1.038(0.027)
r2808-c2-n54	227.755(8.307)	20.162(0.752)	10.592(0.480)	11.587(0.279)	8.028(1.016)	0.703(0.023)
r1010-c3-n65	225.923(10.506)	19.969(0.914)	10.705(0.445)	11.584(0.278)	8.461(1.304)	0.612(0.022)
r3734-c4-n22	227.332(5.318)	19.843(0.472)	10.892(0.474)	11.935(0.206)	9.101(0.795)	0.910(0.027)
r2923-c5-n42	225.524(6.732)	19.855(0.604)	10.792(0.491)	11.668(0.290)	8.577(1.146)	0.803(0.031)
r64-c6-n383	206.626(12.186)	18.237(1.147)	10.731(0.601)	11.484(0.251)	10.244(0.525)	-0.233(0.027)
r3-c7-n700	197.011(13.330)	17.304(1.262)	10.974(0.421)	11.450(0.233)	10.403(0.493)	-0.136(0.031)
r0-c8-n450	203.102(12.104)	17.849(1.168)	11.005(0.364)	11.491(0.276)	9.993(0.762)	0.074(0.028)
r2-c9-n724	196.587(12.436)	17.266(1.197)	11.011(0.330)	11.438(0.255)	10.352(0.546)	-0.034(0.034)
r7-c10-n679	210.348(12.248)	18.536(1.154)	10.949(0.324)	11.521(0.241)	9.469(0.937)	0.203(0.049)
r157-c11-n295	217.587(12.293)	19.196(1.130)	10.913(0.364)	11.542(0.259)	9.154(0.894)	0.366(0.037)
r724-c12-n199	223.904(11.864)	19.829(1.053)	10.776(0.316)	11.522(0.227)	8.582(1.015)	0.492(0.041)
r1332-c13-n85	213.168(7.387)	18.719(0.671)	11.045(0.426)	11.492(0.220)	10.353(0.503)	-0.878(0.047)
r1896-c14-n29	211.066(5.828)	18.389(0.531)	11.308(0.449)	11.448(0.221)	10.734(0.307)	-1.003(0.024)
r2243-c15-n21	213.168(5.438)	18.415(0.275)	11.556(0.552)	11.497(0.215)	10.805(0.169)	-1.108(0.040)
r123-c16-n188	212.108(12.498)	18.694(1.123)	10.741(0.401)	11.484(0.311)	10.455(0.507)	-0.629(0.023)
r608-c17-n171	213.243(9.660)	18.777(0.883)	10.837(0.471)	11.484(0.288)	10.440(0.449)	-0.725(0.032)
r61-c18-n434	207.572(13.188)	18.312(1.208)	10.755(0.480)	11.468(0.280)	10.307(0.545)	-0.346(0.038)
r87-c19-n460	211.186(10.993)	18.602(0.935)	10.710(0.492)	11.522(0.397)	10.336(0.522)	-0.495(0.048)
r4591	222.273	19.600	10.633	11.782	8.464	1.015
r2808	208.451	18.274	11.405	11.418	9.926	0.709
r1010	193.387	17.169	10.895	11.275	9.885	0.614
r3734	215.074	18.936	10.897	11.681	9.831	0.886
r2923	209.240	17.940	11.993	11.500	10.388	0.786
r64	177.599	15.682	11.185	11.059	10.454	-0.240
r3	162.513	14.187	11.162	11.409	10.351	-0.109
r0	153.216	12.385	12.328	11.166	11.522	0.072
r2	161.329	13.640	11.646	11.520	11.011	-0.081
r7	166.053	14.219	11.931	11.058	10.563	0.152
r157	181.369	15.899	11.117	11.491	9.894	0.350
r724	190.117	16.732	10.709	11.644	10.208	0.432
r1332	196.530	16.629	11.925	11.699	10.835	-0.818
r1896	201.307	17.689	10.848	11.445	10.894	-1.014
r2243	204.053	18.058	10.706	11.298	10.845	-1.067
r123	180.065	15.615	11.263	11.577	10.603	-0.639
r608	188.722	16.450	10.992	11.656	10.661	-0.719
r61	177.471	15.416	11.214	11.459	10.792	-0.311
r87	178.455	15.590	11.329	11.108	10.858	-0.434

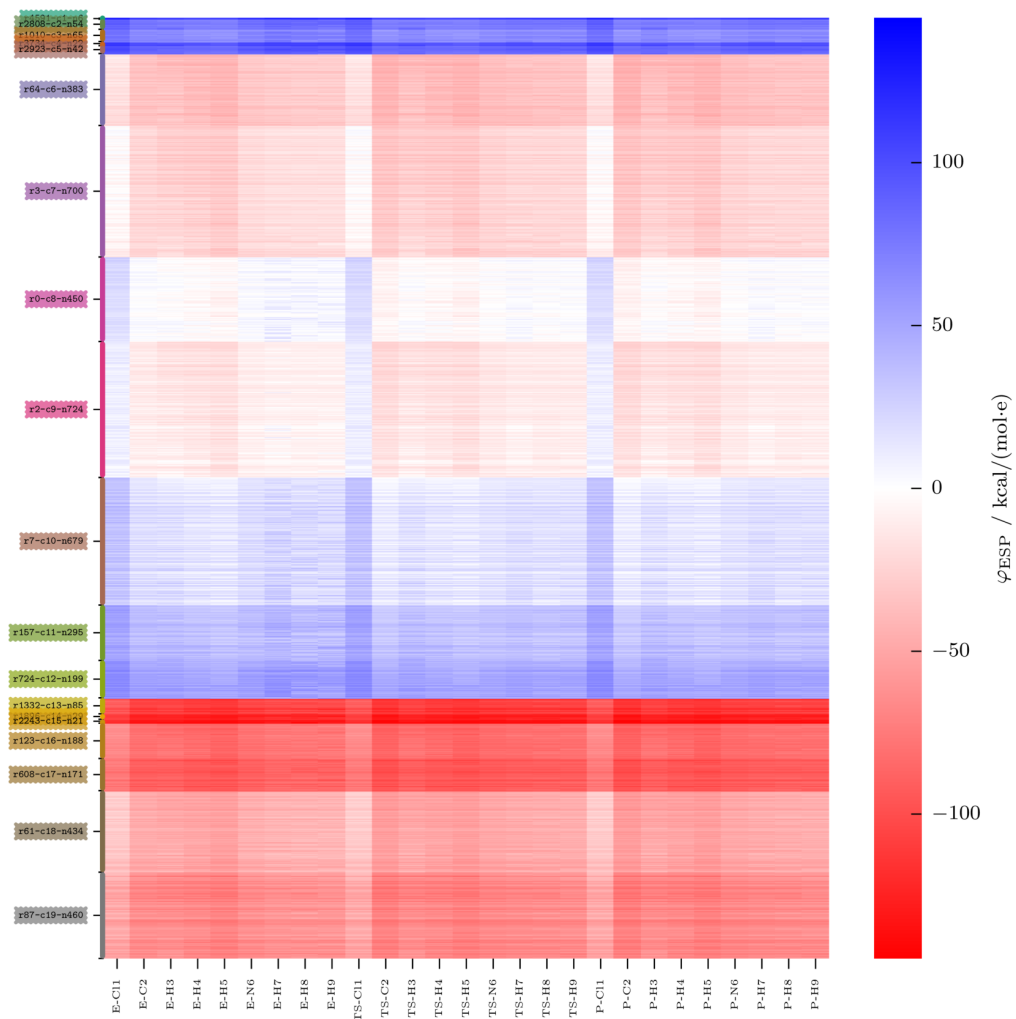


Fig. S28: $N_{\text{Ch}} = 10$ GOCATs (PM7): Heatmap of the complete database chunked into 19 clusters; for illustration details, see Fig. S6.

S5 PM7: $N_{\text{Ch}} = 10$ case (summed charge neutrality)

S5.1 Cluster Analysis

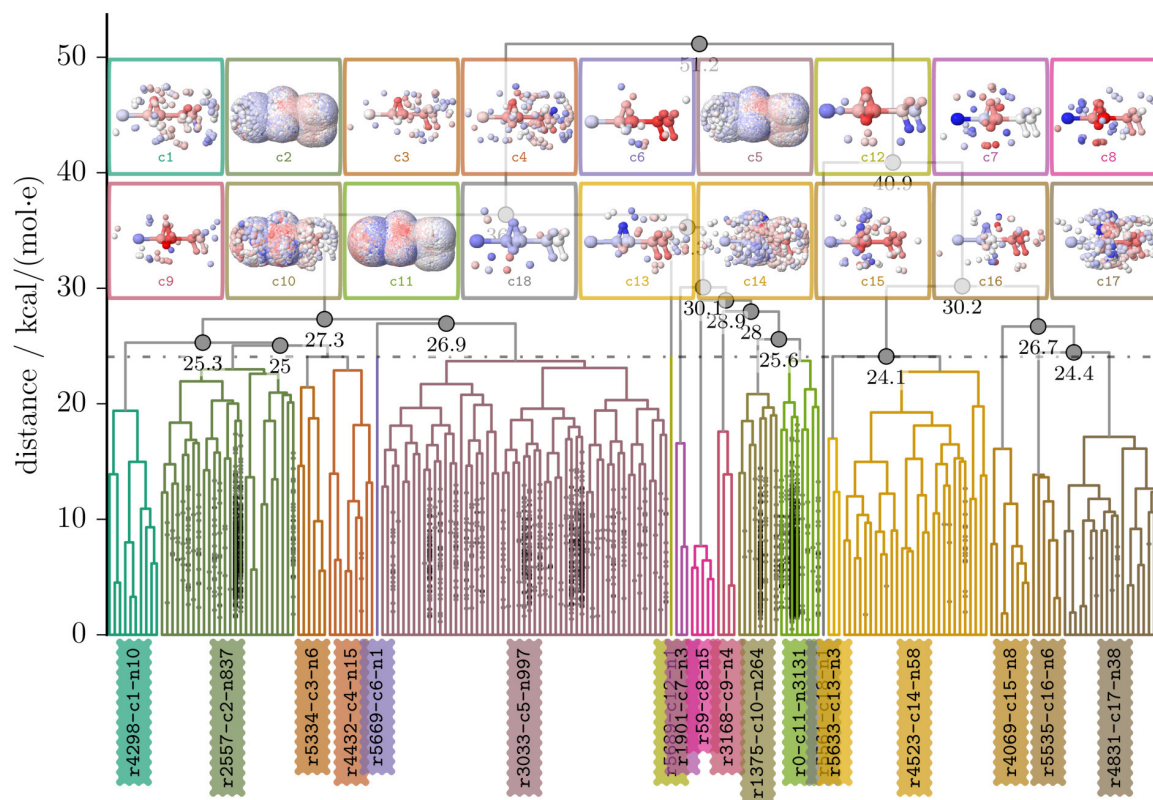


Fig. S29: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral): Dendrogram of final database with 5388 non-identical individuals using the average linkage strategy, cut into 18 different clusters, (below) $24.06 \text{ kcal mol}^{-1} e^{-1}$ (dotted line); for illustration details, see Fig. S12.

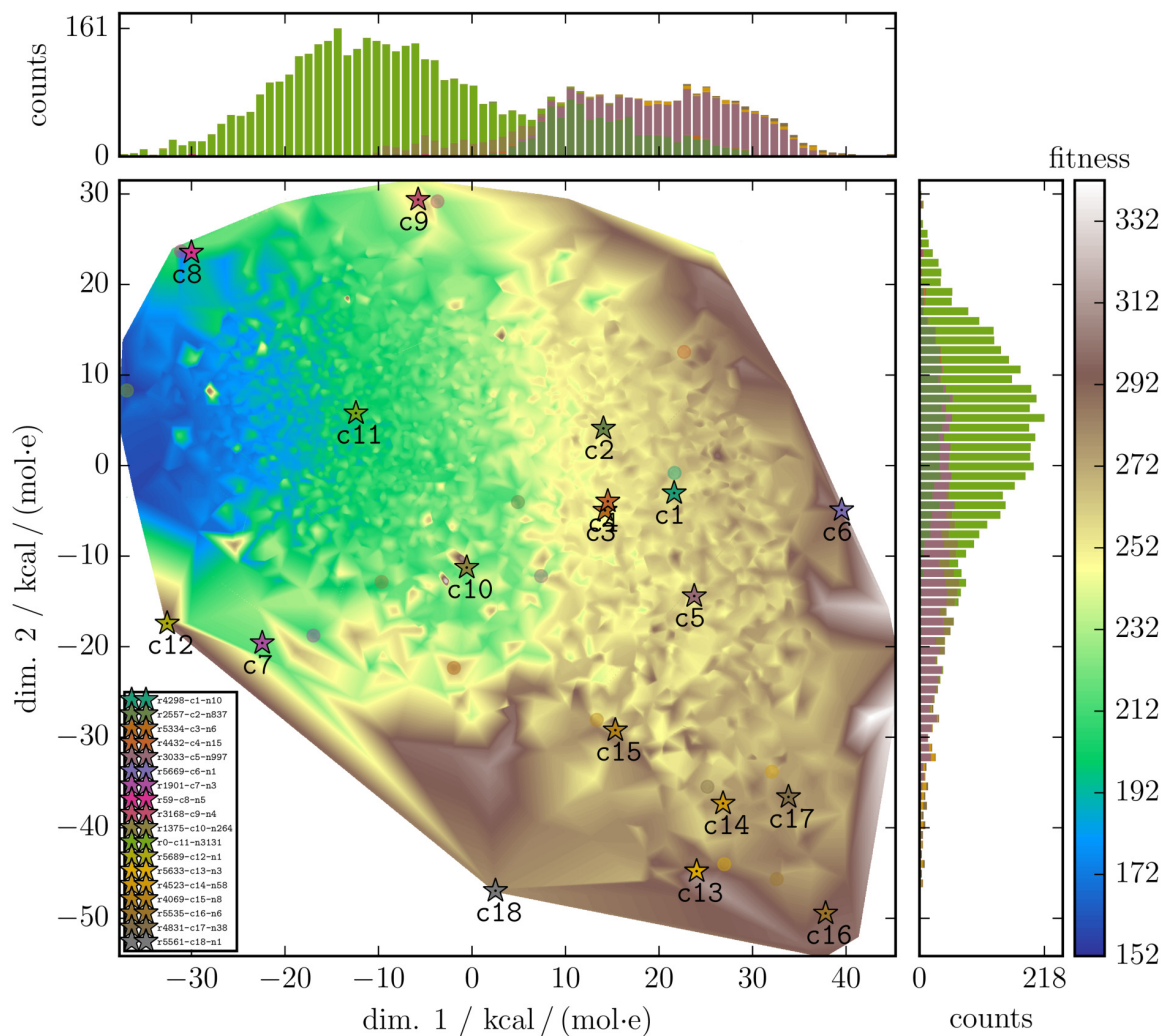


Fig. S30: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral): Multidimensional Scaling as 2D projection similar to Fig. 6 (main article); for illustration details, see Fig. S1. With this zoom into the neutral domain of the (almost quite flexible) $N_{\text{Ch}} = 10$ case, we can observe an overly rugged surface with a clear slope to the “upper left” in this plot. The by far biggest cluster c11 is also the one with the best individuals in it, while the (very symmetric) special best rank(s), e.g. r0 shown in the main article, is again at the edge of the distribution (see histograms).

S5.2 Reaction Paths (Selected Clusters)

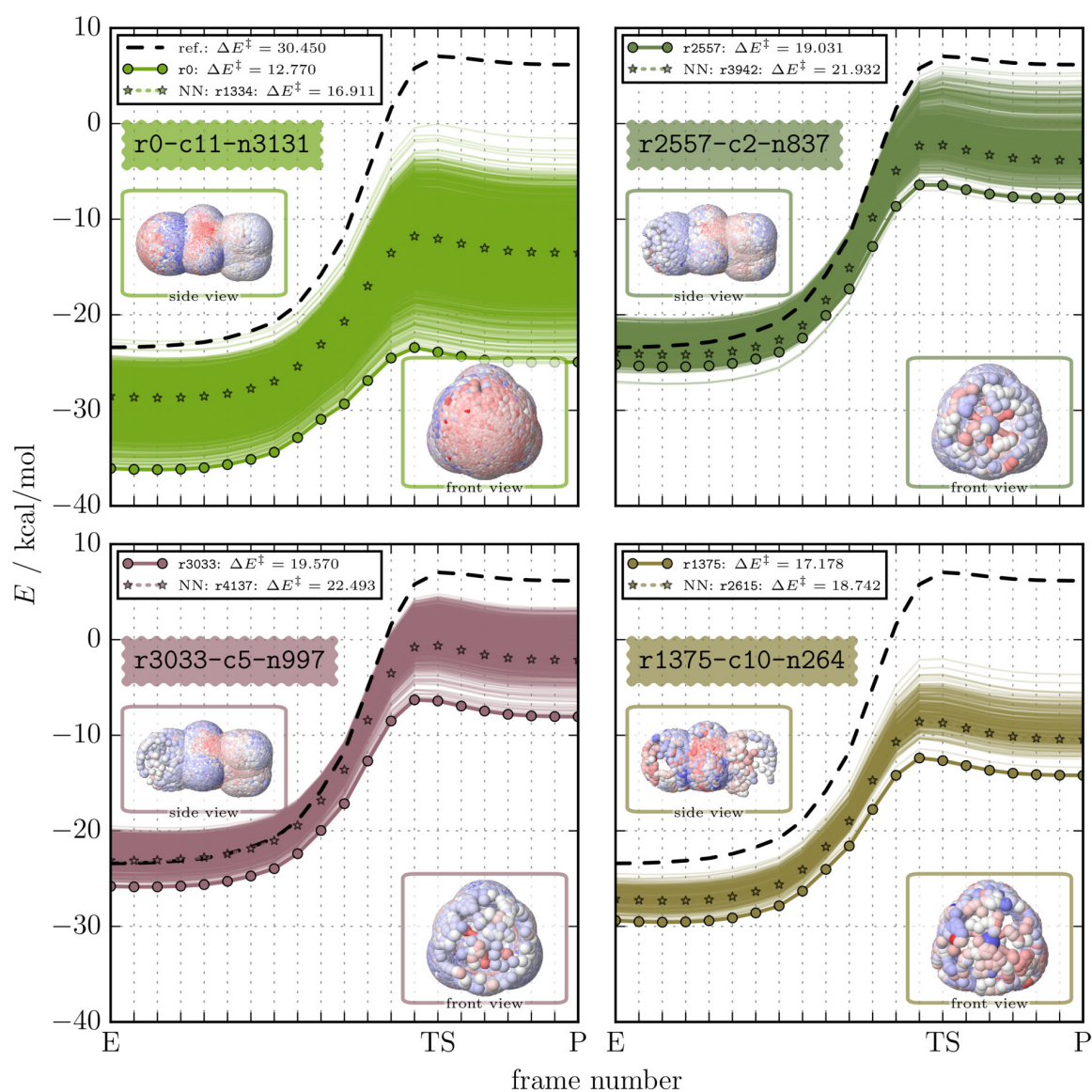


Fig. S31: $N_{Ch} = 10$ GOCATs (PM7, neutral): Reaction paths of the dominant big GOCAT cluster (c11) and other 3 bigger ones c2, c5, c10. For illustration details, see Fig. S3. The average ESP values discussed in the main article are very close to the given nearest neighbor (NN) path in c11, i.e., of r1134 with $\Delta E^\ddagger = 16.911$ kcal mol⁻¹.

S5.3 Selected Details

Table S4: Mean and standard deviations (parentheses) of the electrostatic potentials (φ_{ESP}) at the 3 core frames (E, TS, P) of the Menshutkin reaction for different cases.

atom name ^a	$\varphi_{\text{ESP}} / \text{kcal mol}^{-1} \text{e}^{-1}$			
	gas phase GOCATs ^b	COSMO GOCATs ^c	pure COSMO ^d	DFT GOCATs ^e
E-C11	13.70(3.67)	39.31(2.69)	13.95	5.99(3.83)
E-C2	-5.07(1.54)	18.90(3.30)	-2.83	-3.71(4.28)
E-H3	-9.96(2.52)	10.28(4.15)	-13.16	-1.83(6.61)
E-H4	-12.90(2.54)	12.37(4.05)	-6.04	-16.41(6.22)
E-H5	-7.68(2.82)	16.94(5.05)	-3.12	-4.31(6.99)
E-N6	-4.12(1.43)	-53.31(3.40)	-7.38	-1.91(2.88)
E-H7	-2.44(2.65)	-54.73(4.09)	-19.25	-4.16(4.12)
E-H8	-2.36(2.69)	-59.52(4.49)	-14.42	1.49(7.03)
E-H9	-3.21(2.07)	-54.31(3.95)	-19.02	3.44(7.99)
TS-C11	14.12(3.86)	37.90(2.65)	38.39	5.95(3.93)
TS-C2	-14.54(2.00)	10.55(3.52)	-4.17	-9.54(3.66)
TS-H3	-11.00(2.61)	8.08(4.12)	-8.52	-3.38(6.93)
TS-H4	-13.93(2.45)	7.45(4.04)	-8.33	-19.22(6.35)
TS-H5	-8.70(3.03)	11.92(4.98)	-8.44	-5.64(7.27)
TS-N6	-7.28(1.86)	-48.09(3.27)	-37.21	-5.88(2.73)
TS-H7	-5.66(1.98)	-54.01(3.92)	-42.46	-8.52(4.84)
TS-H8	-5.46(1.96)	-43.12(3.33)	-42.32	-2.71(5.99)
TS-H9	-5.96(1.86)	-53.42(3.87)	-42.51	0.12(6.43)
P-C11	13.79(3.71)	42.54(2.67)	85.37	5.97(3.87)
P-C2	-15.28(2.05)	-13.49(3.06)	-33.45	-10.09(3.62)
P-H3	-11.42(2.61)	0.93(3.88)	-19.16	-3.78(6.86)
P-H4	-14.43(2.43)	-3.11(4.42)	-19.70	-19.31(6.23)
P-H5	-9.14(3.04)	0.15(4.66)	-20.66	-5.88(7.18)
P-N6	-7.13(1.84)	-50.53(3.38)	-82.85	-5.97(2.74)
P-H7	-5.46(1.95)	-56.25(4.00)	-92.43	-8.57(4.86)
P-H8	-5.24(1.93)	-44.00(3.36)	-92.92	-2.76(5.99)
P-H9	-5.77(1.83)	-55.67(3.95)	-92.80	0.09(6.41)

^a E, TS and P atoms given. H3-5 are the ones at C2, H7-9 are the ones connected to N6.

^b $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, gas phase NEB path) described in the main article. Statistics of r0-c11-n3131 of corresponding database are presented.

^c $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, COSMO NEB path) described in the main article as stabilization of COSMO (H_2O) structures within GOCATs. Statistics of r0-c16-n2210 of corresponding database are presented.

^d Calculated electrostatic potentials at the core atoms in a COSMO (H_2O) calculation.

^e $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral, gas phase NEB path). Statistics of r0-c10-n695 of corresponding database are presented.

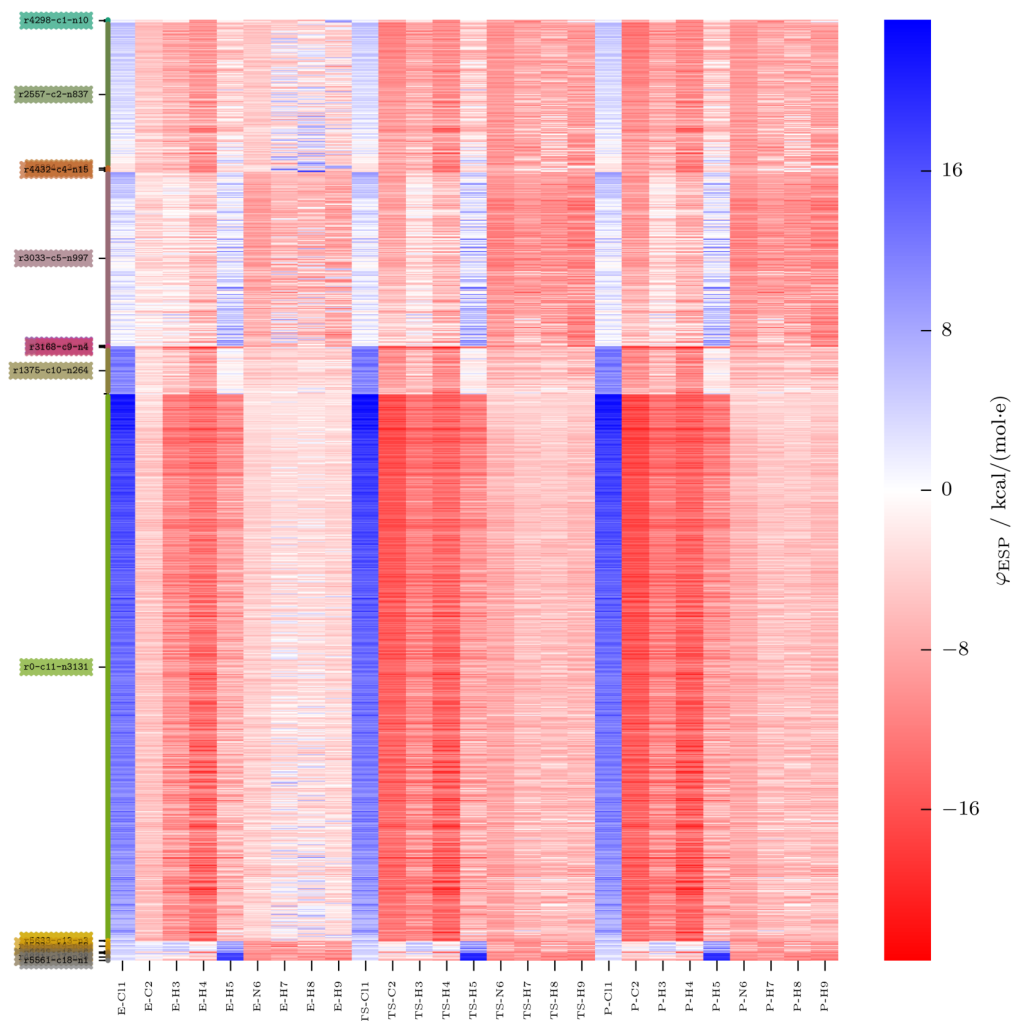


Fig. S32: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral): Heatmap of the complete database chunked into 18 clusters; for illustration details, see Fig. S6.

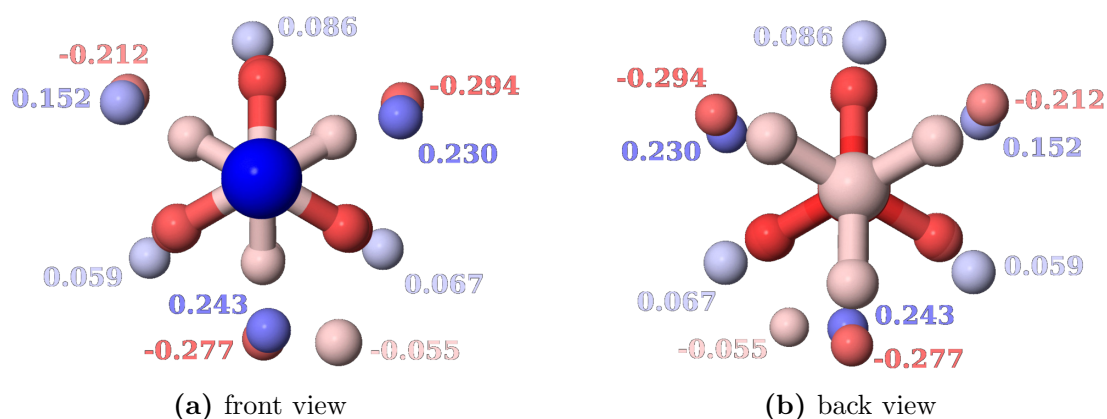


Fig. S33: Single $N_{\text{Ch}} = 10$ GOCAT individual (PM7, neutral): two different views of r_0 (of c_{11}) with values given for the partial charges. Both partial charges and the atoms of the selected core frames (E, TS, P) are colored red/blue in the ranges $[-0.537, +0.537] e$ for charges and $[-23.578, +23.578] \text{ kcal mol}^{-1} e^{-1}$ for ESP values. Complementary to Fig. 8 (main article).

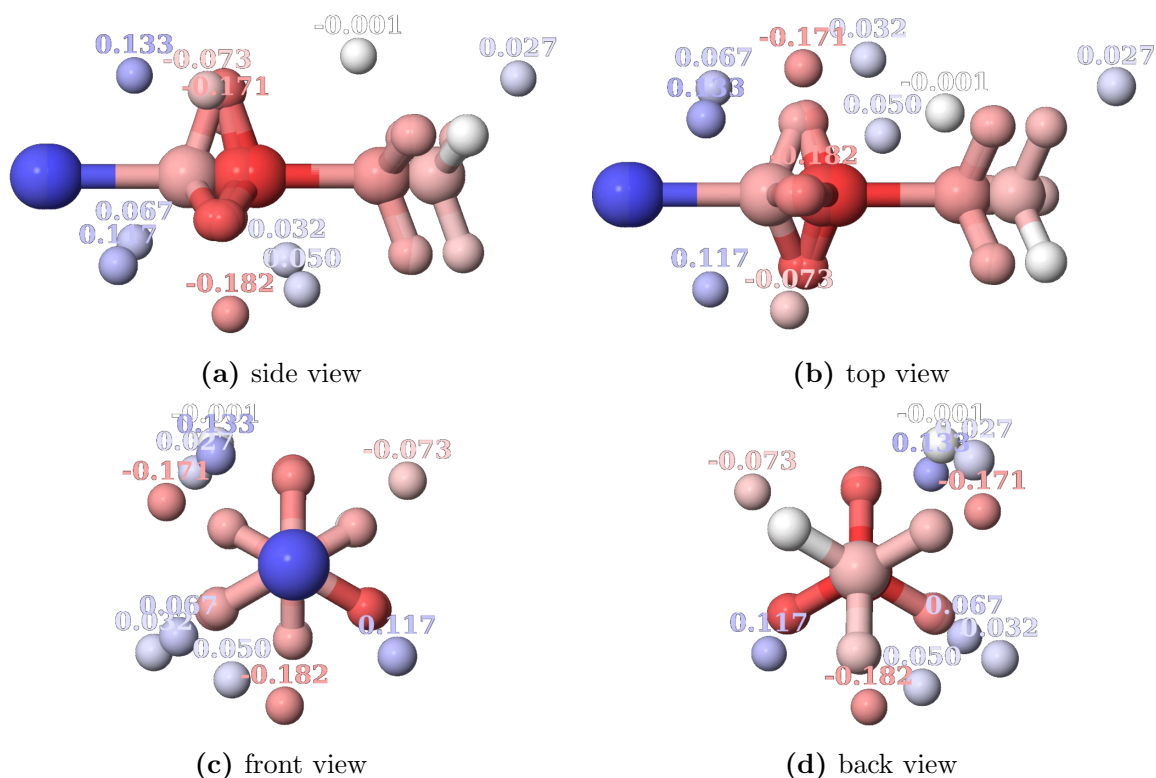


Fig. S34: Single GOCAT individual (PM7, neutral): four different views of the nearest neighbor GOCAT to the calculated cluster center of c_{11} : r_{1334} for the $N_{\text{Ch}} = 10$ case. Values given for the partial charges. Both partial charges and the atoms of the core frames (E, TS, P shown) are colored red/blue in the ranges $[-0.537, +0.537] e$ for charges and $[-23.578, +23.578] \text{ kcal mol}^{-1} e^{-1}$ for ESP values, explicitly: Cl: 14.17, 14.51, 14.25; C: $-5.66, -15.66, -16.34$; N: $-4.57, -8.03, -7.87$.

S6 PM7: $N_{\text{Ch}} = 10$ case (summed charge neutrality and no TS gradient norm threshold)

S6.1 Complementary Illustration

Section missing in the main article: In all other cases, usually gradient norm thresholds were included in the objective function to still have (near) stationary points on the effective new PES within the GOCAT. For the educt and product frames, this is strictly necessary, but for the TS frame, a finally found *minimal* barrier should include already a TS point that is the lowest possible (1. order) saddle point to the next valley, *within* the pre-fixed set of frames excluding E and P. Thus, here we have the exact same setting like the Section(s) before (Section S5) and described in the “Methods” Section 2 of the main article or in Section S1 above, except for the gradient threshold for a fixed TS frame. Before, more or less “vertical” energetical shifts between a pristine PES and the GOCAT PES were generated, now we also have a possibility to let the TS shift discretely.

Like demonstrated in the main article, we already have big barrier decreases for the “vertical” mode. Using a complete other path (with outwards rotation of NH_3) like the COSMO path (next Section), we have exothermicity, shifts of the TS to the E frame and bigger barrier decreases. The case of this additional Section without TS gradient norms in the objective function now illustrates an “intermediate” case between the two extremes. Still, this is in accordance with the main discussion line of the article. At any rate, in future applications the first and most important change is the re-optimization/relaxation of a minimum energy path for each GOCAT separately, which would make the decision about “vertical” mode or less restrictions obsolete.

S6.2 Cluster Analysis

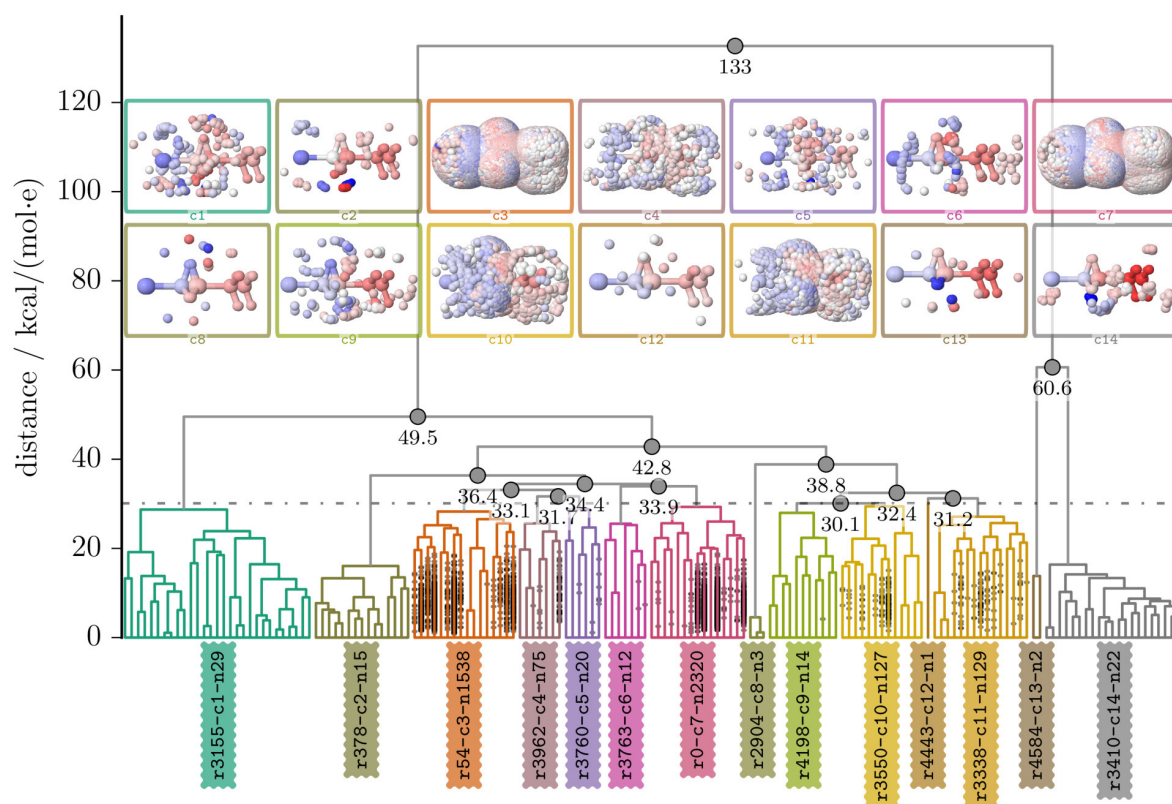


Fig. S35: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, no TS fixation): Dendrogram of final database with 4307 non-identical individuals using the average linkage strategy, cut into 14 different clusters, (below) $30.11 \text{ kcal mol}^{-1} \text{e}^{-1}$ (dotted line); for illustration details, see Fig. S12.

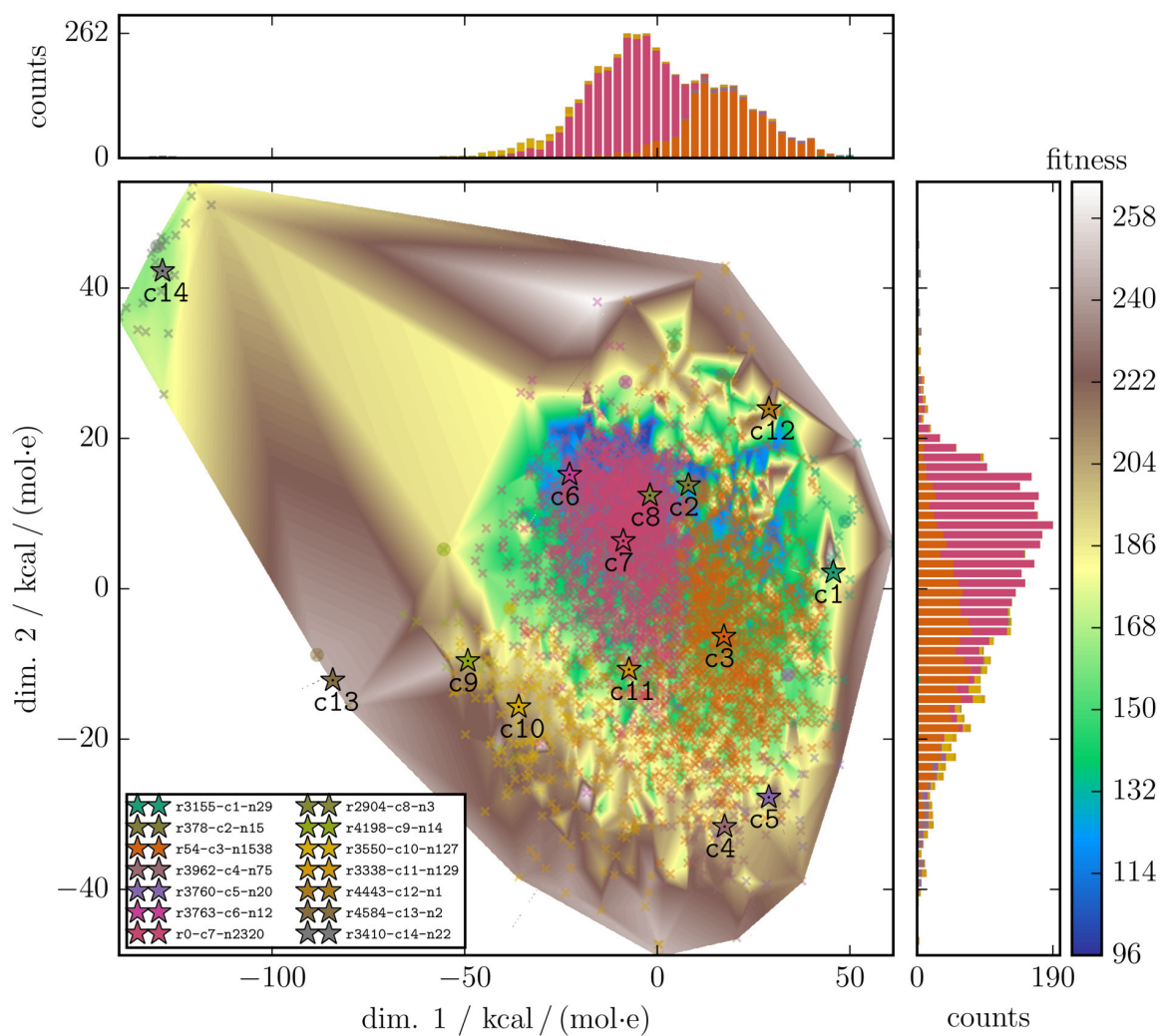


Fig. S36: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, no TS fixation): Multidimensional Scaling as 2D projection of the higher dimensional ESP-distance data; for illustration details, see Fig. S13.

S6.3 Reaction Paths (Selected Clusters)

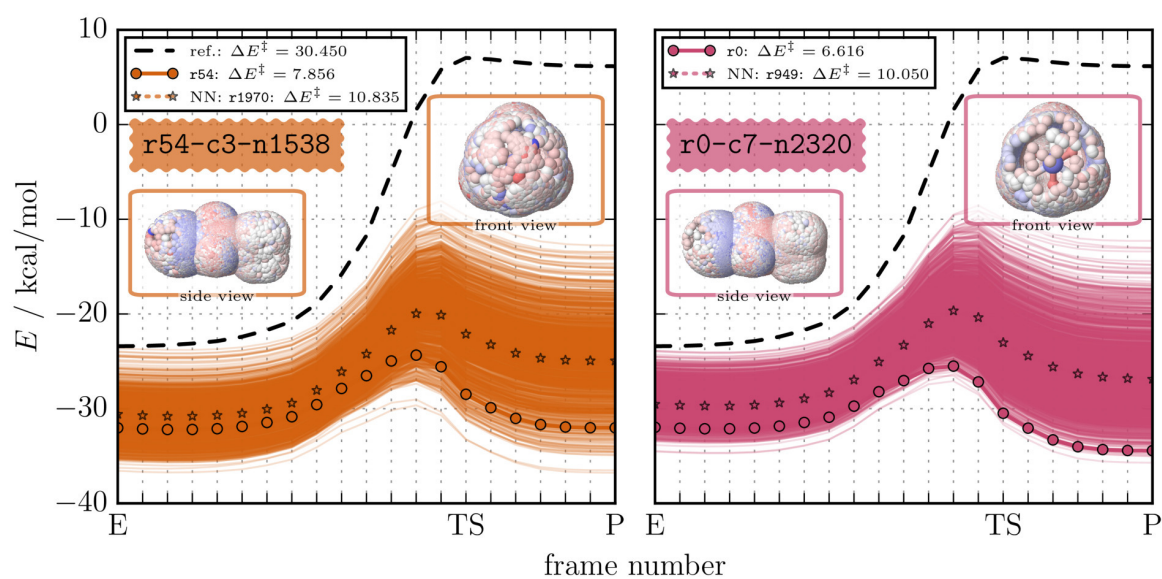


Fig. S37: $N_{\text{Ch}} = 1$ GOCATs (PM7, neutral, no TS fixation): Reaction paths of the two biggest GOCAT clusters: c3 and c7. For illustration details, see Fig. S3. Note: The TS indicated at the x -axis is that of the gas-phase path. Now the highest energy frame (TS) in each GOCAT is shifted to the E frame and shows up also a local minimum of gradient norm present except for E and P.

S6.4 Selected Details

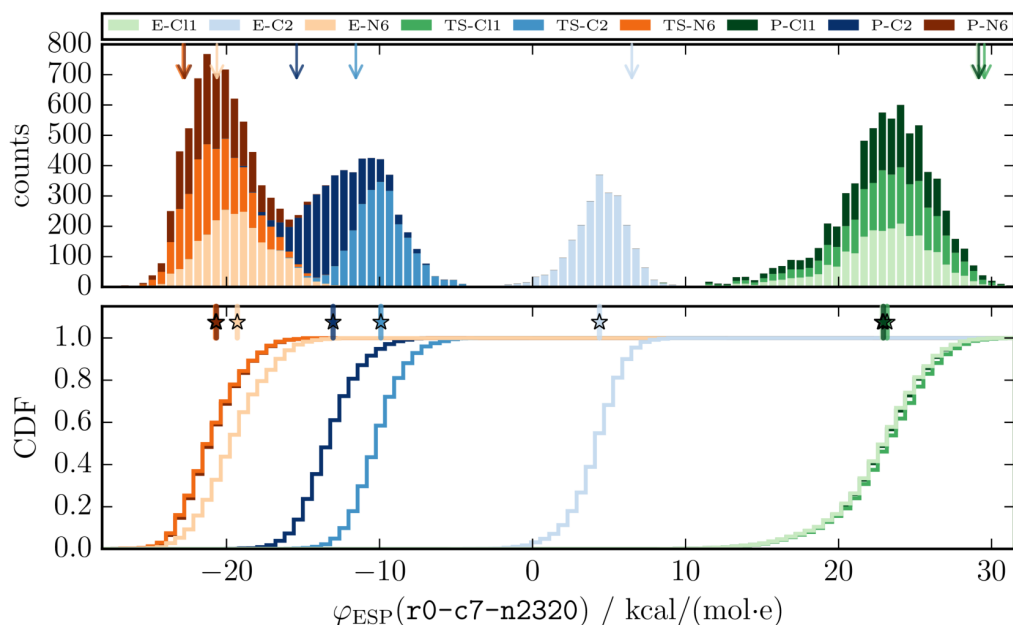


Fig. S38: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, no TS fixation): Stacked histogram (top) and cumulative distribution functions (bottom) of φ_{ESP} for the cluster surrounding the best GOCAT: c7. Arrows indicate the explicit φ_{ESP} values of the best rank within their distributions, r0 (given in Fig. S39), following the coloring of the 9 separate atoms shown in the legend. Below: Cumulative distribution function showing the spread or (if present) skewness; vertical bars with stars are the computed *average* φ_{ESP} values of that cluster at the corresponding atoms. All average φ_{ESP} values in $\text{kcal mol}^{-1}\text{e}^{-1}$ for E, TS and P frames (standard deviation in parentheses): Cl: 21.30(3.17), 21.71(3.30), 21.39(3.19); C: 1.49(2.27), -12.06(3.11), -14.48(3.30); N: -13.06(2.59), -16.31(3.28), -16.17(3.25). As general trend of φ_{ESP} (compare with both histograms in the main article Figs. 7 and 10): The total φ_{ESP} range increases and the N atom is already more negatively embedded than in Fig. 7 (with TS fixation), showing the trend to reach embeddings as in Fig. 10 (COSMO case).

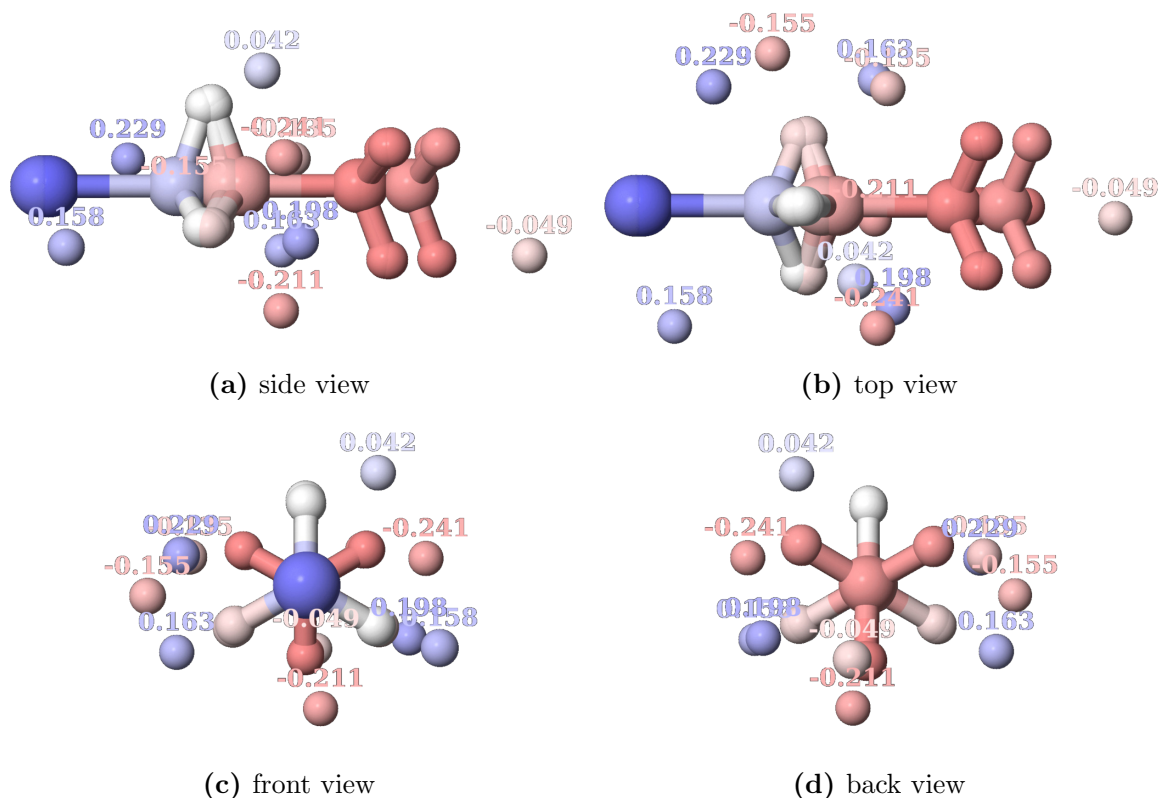


Fig. S39: Single GOCAT individual: four different views of the best rank found during the GA runs without checking the TS gradient norm (i.e., some discrete shifts possible) of cluster *c7*: *r0* for $N_{\text{Ch}} = 10$. Values given for the partial charges. Both partial charges and the atoms of the core frames (E, TS, P shown) are colored red/blue in the ranges $[-0.847, +0.847]e$ for charges and $[-60.217, +60.217] \text{ kcal mol}^{-1} e^{-1}$ for ESP values, explicitly: Cl: 29.06, 29.53, 29.17; C: 6.49, -11.55 , -15.42 ; N: -20.63 , -22.86 , -22.74 .

S7 PM7: $N_{\text{Ch}} = 10$ stabilizing COSMO path (summed charge neutrality)

S7.1 Cluster Analysis

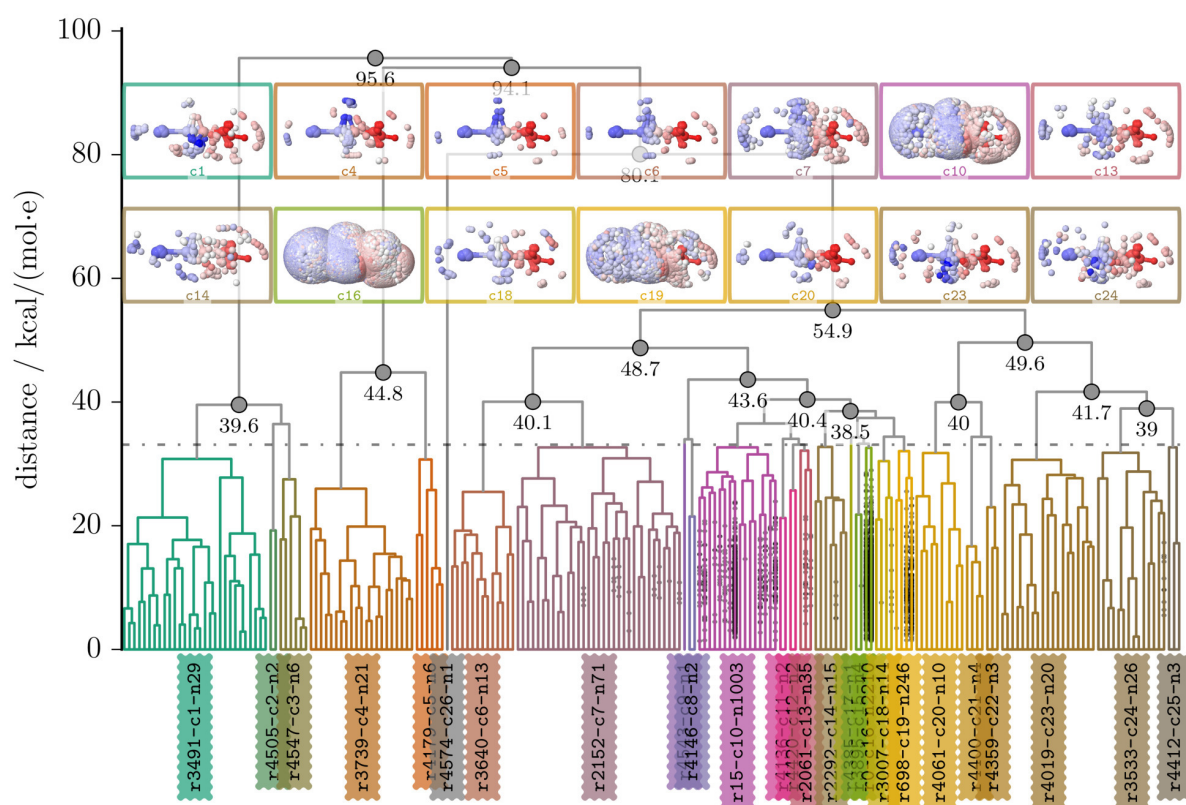


Fig. S40: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, COSMO): Dendrogram of final database with 3749 non-identical individuals using the average linkage strategy, cut into 26 different clusters, (below) $33.11 \text{ kcal mol}^{-1} \text{e}^{-1}$ (dotted line); for illustration details, see Fig. S12.

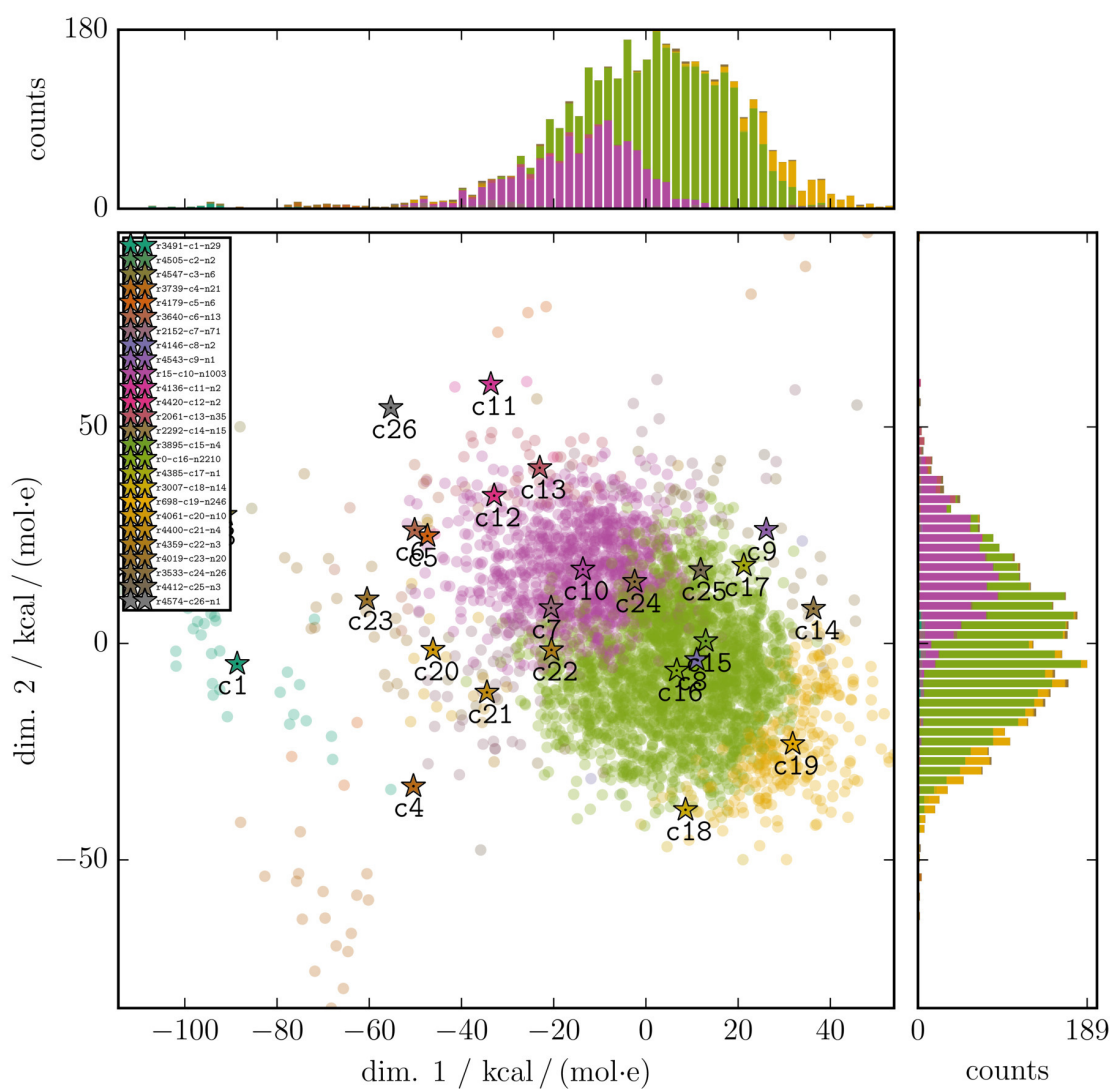


Fig. S41: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, COSMO): Multidimensional Scaling as 2D projection of the higher dimensional ESP-distance data; for illustration details, see Fig. S13.

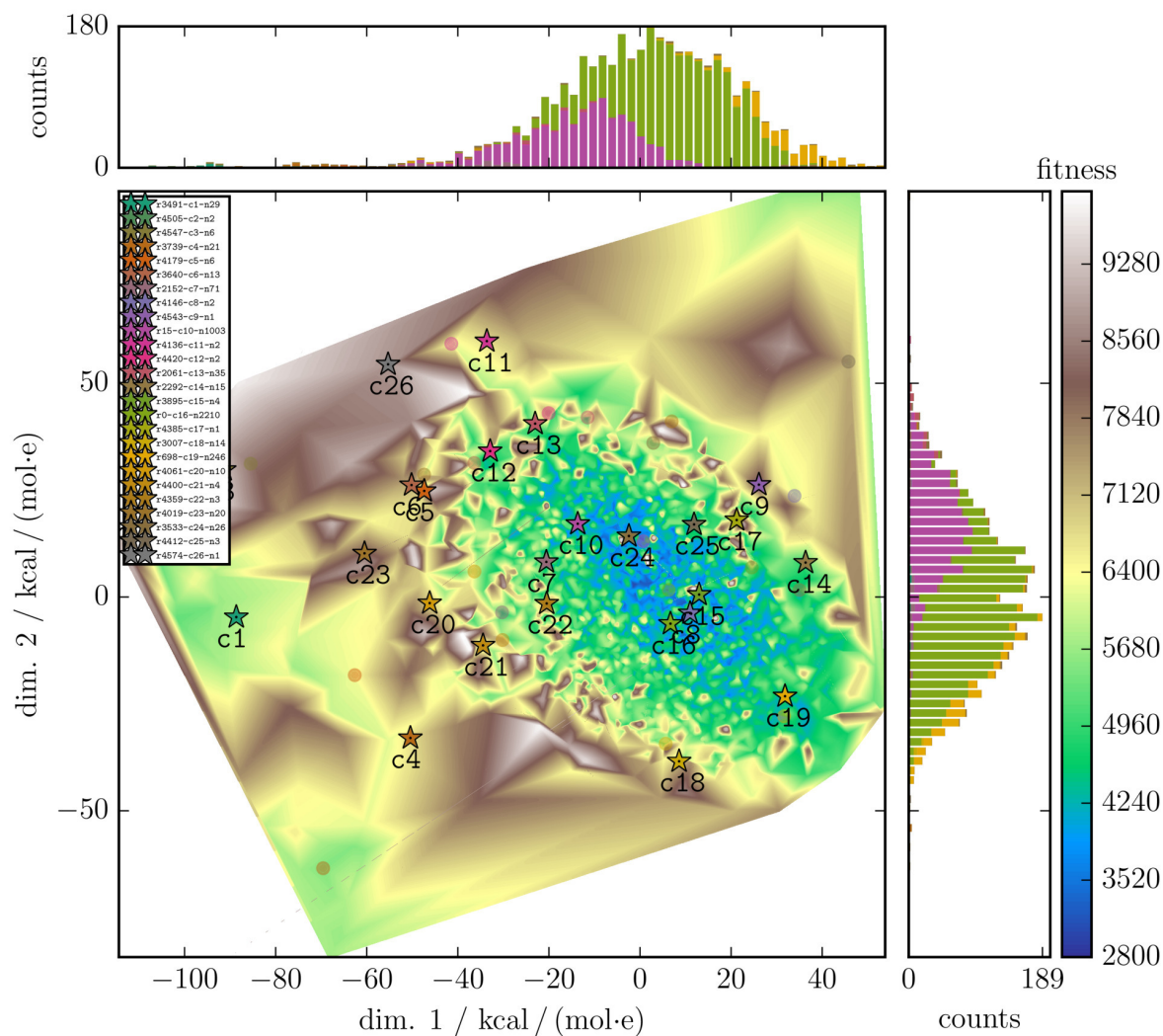


Fig. S42: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, COSMO): Multidimensional Scaling as 2D projection including fitness values similar to Fig. S41; for illustration details, see Fig. S1. Notice that within this objective function, the gradient norms on the stationary points are the most dominant part and essentially mirror the fitness surface shown here. The most dominant cluster (c11) and its best rank is embedded in the center of the overall distribution and does not sit at the edge of the clusters as it does in Section S5.

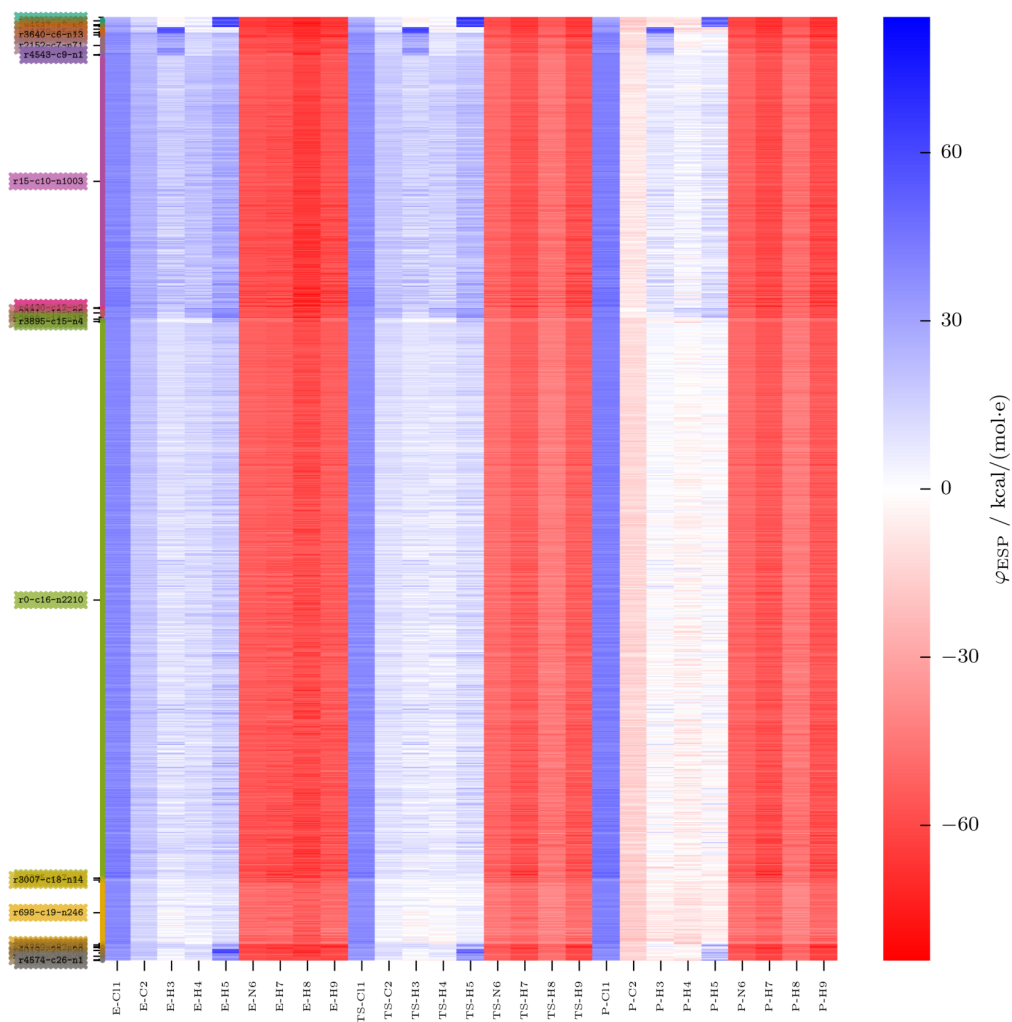


Fig. S43: $N_{\text{Ch}} = 10$ GOCATs (PM7, neutral, COSMO): Heatmap of the complete database chunked into 26 clusters; for illustration details, see Fig. S6. Starting with the COSMO reaction frames, there is a very clear trend (i.e., most of the solutions are very similar at least with respect to the dominant absolute values) in all the φ_{ESP} values in order to stabilize that path.

S7.2 Selected Details

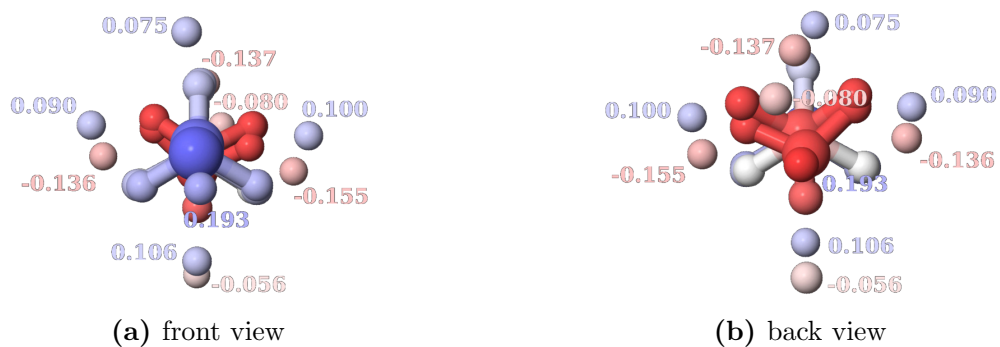


Fig. S44: Single $N_{\text{Ch}} = 10$ GOCAT individual (PM7, neutral, COSMO): r0 (of c16) case with values given for the partial charges. Both partial charges and the atoms of the core frames (E, TS, P shown) are colored red/blue in the ranges $[-0.751, +0.751]$ e for charges and $[-84.142, +84.142]$ kcal mol $^{-1}$ e $^{-1}$ for ESP values. Complementary to Fig. 11 (main article).

S8 DFT: $N_{\text{Ch}} = 10$ case (summed charge neutrality)

S8.1 Cluster Analysis

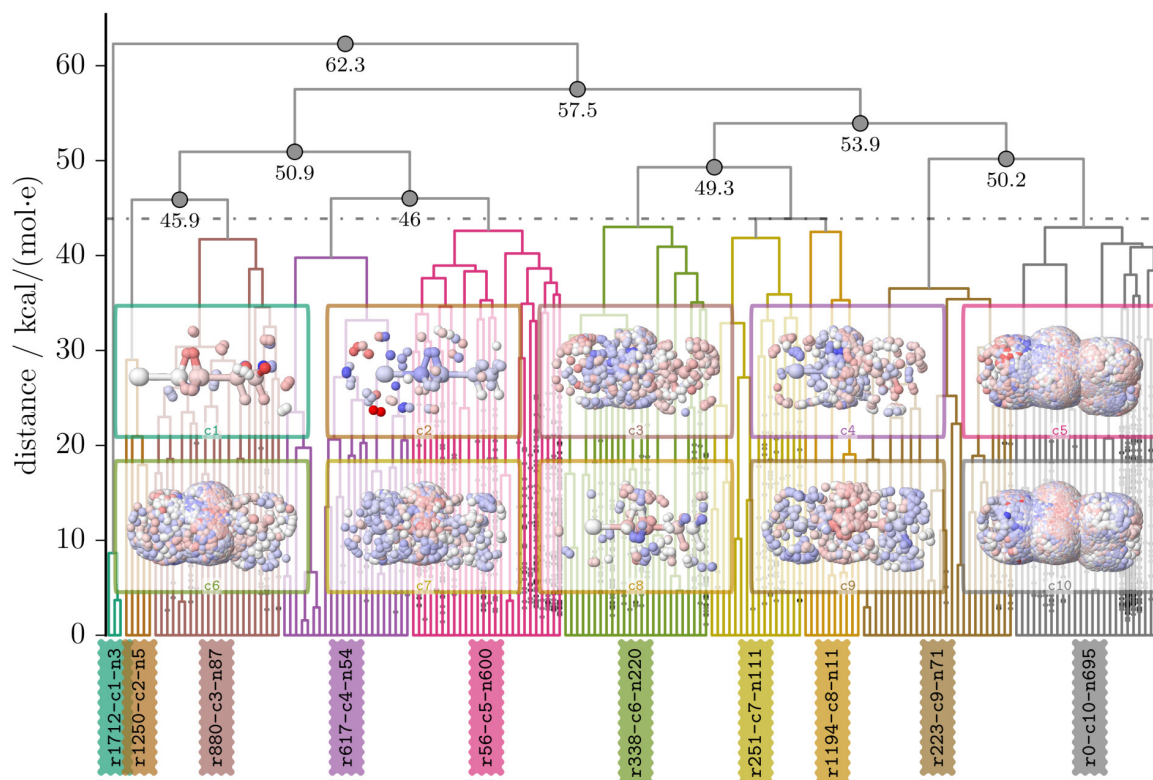


Fig. S45: $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral): Dendrogram of final database with 1856 non-identical individuals using the average linkage strategy, cut into 10 different clusters, (below) $43.89 \text{ kcal mol}^{-1} \text{e}^{-1}$ (dotted line); for illustration details, see Fig. S12.

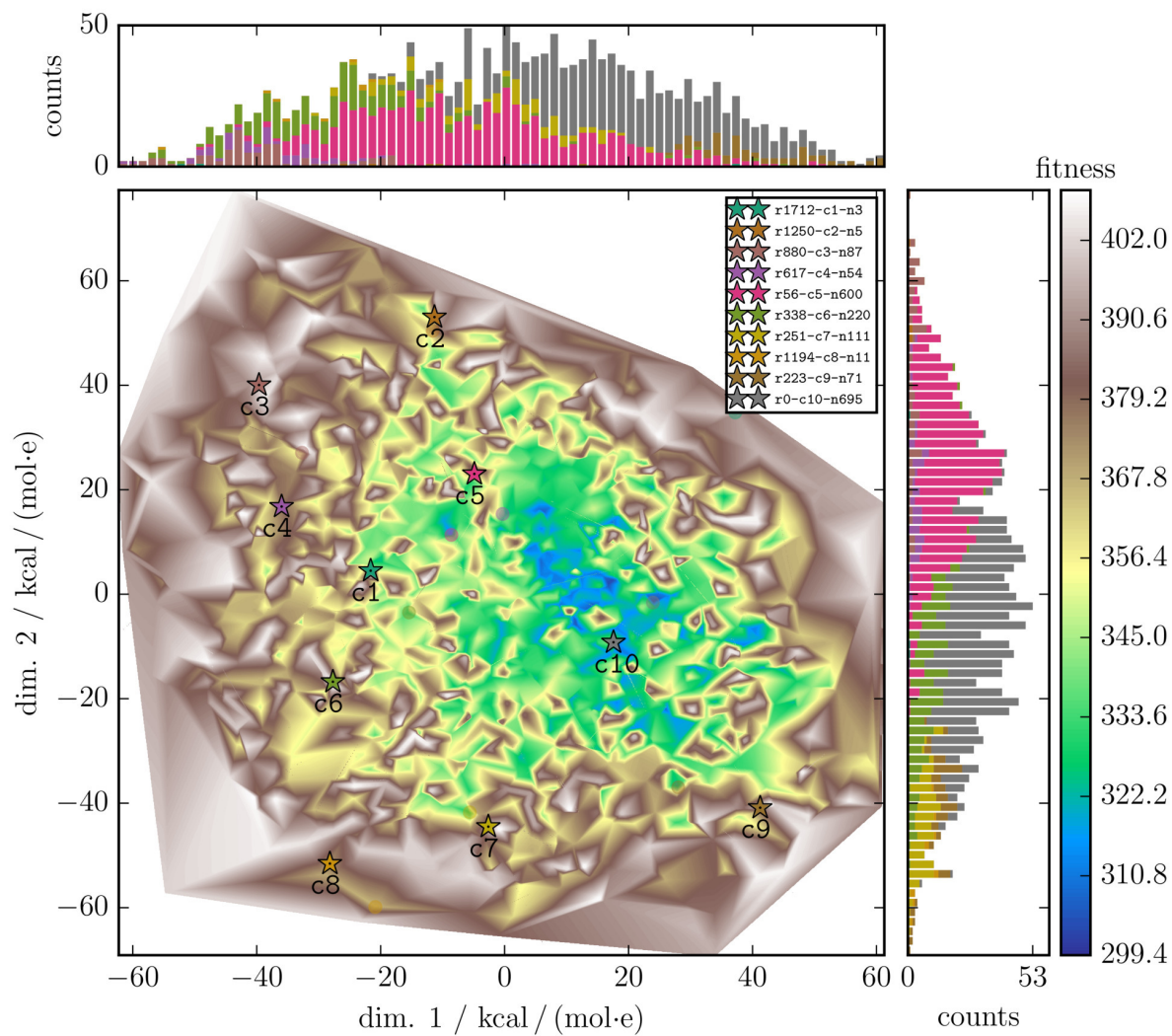


Fig. S46: $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral): Multidimensional Scaling as 2D projection; for illustration details, see Fig. S1.

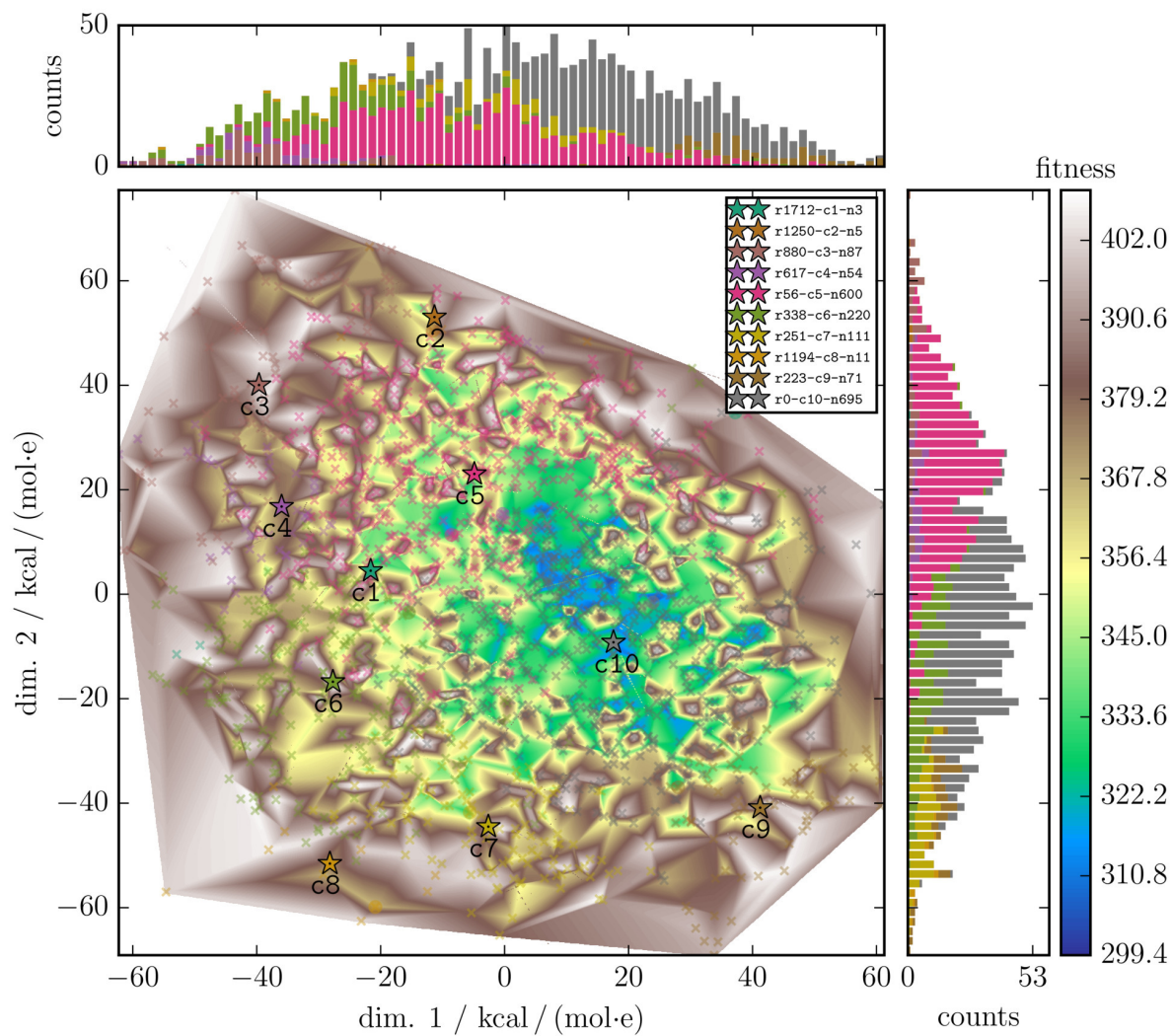


Fig. S47: $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral): Similar plot to Fig. S46. for illustration details, see Fig. S2.

S8.2 Reaction Paths (Selected Clusters)

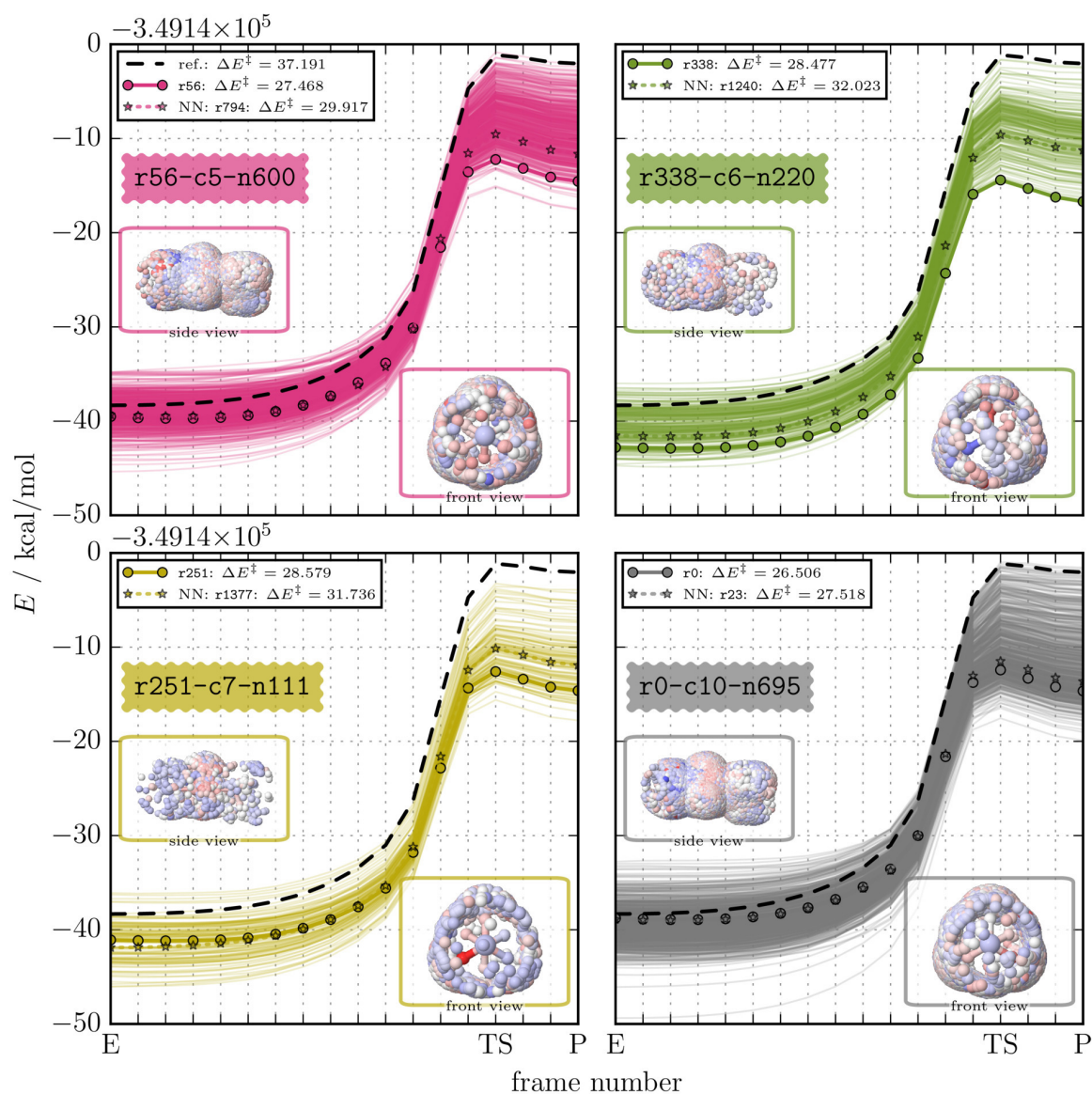


Fig. S48: $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral): Reaction paths of the four biggest GOCAT clusters: c5-c7 and the one with r0: c10. For illustration details, see Fig. S3.

S8.3 Selected Details

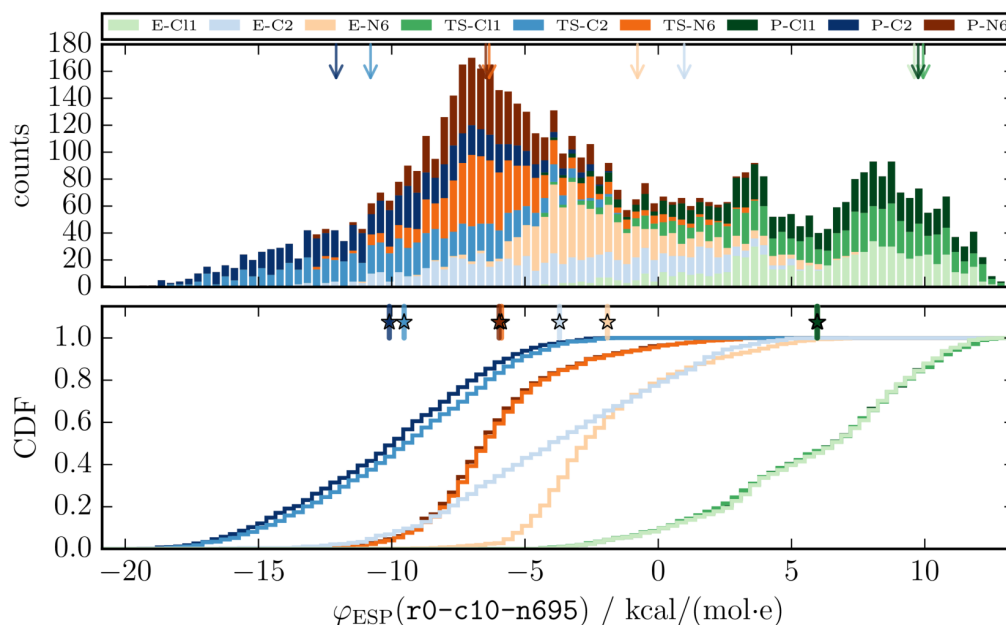


Fig. S49: $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral): Stacked histogram (top) and cumulative distribution functions (bottom) of φ_{ESP} for the cluster surrounding the best GOCAT: c10. Arrows indicate the explicit φ_{ESP} values of the best rank within their distributions, \mathbf{r}_0 (given in Fig. S51), following the coloring of the 9 separate atoms shown in the legend. Below: Cumulative distribution function showing the spread or (if present) skewness; vertical bars with stars are the computed *average* φ_{ESP} values of that cluster at the corresponding atoms. All average φ_{ESP} values in $\text{kcal mol}^{-1} \text{e}^{-1}$ for E, TS and P frames (standard deviation in parentheses): Cl: 5.99(3.83), 5.95(3.93), 5.97(3.87); C: -3.71(4.28), -9.54(3.66), -10.09(3.62); N: -1.91(2.88), -5.88(2.73), -5.97(2.74). These are compiled in Table S4. As most prominent qualitative feature: The trend of φ_{ESP} is the same as discussed in the main article in Section 3.3 (compare with Fig. 7), except for the broader distributions that can be ascribed to the smaller database of fewer GA runs with less iterations. Both in average and in the best rank found, the most negative embedding is at C (with the shift from educt to TS/product), then N and a strong positive Coulomb potential at Cl, in good qualitative agreement with the PM7 results.

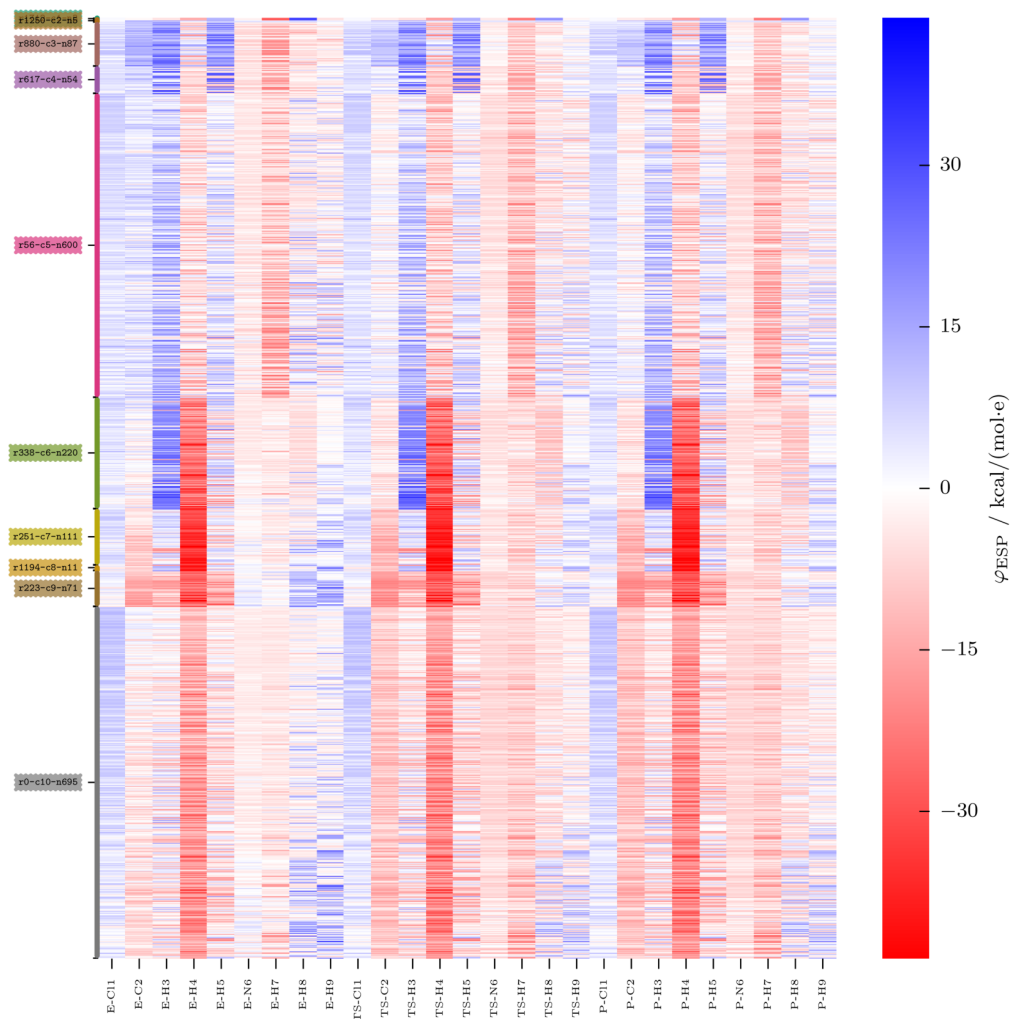


Fig. S50: $N_{\text{Ch}} = 10$ GOCATs (PBE0/def2-TZVP, neutral): Heatmap of the complete database chunked into 10 clusters; for illustration details, see Fig. S6.

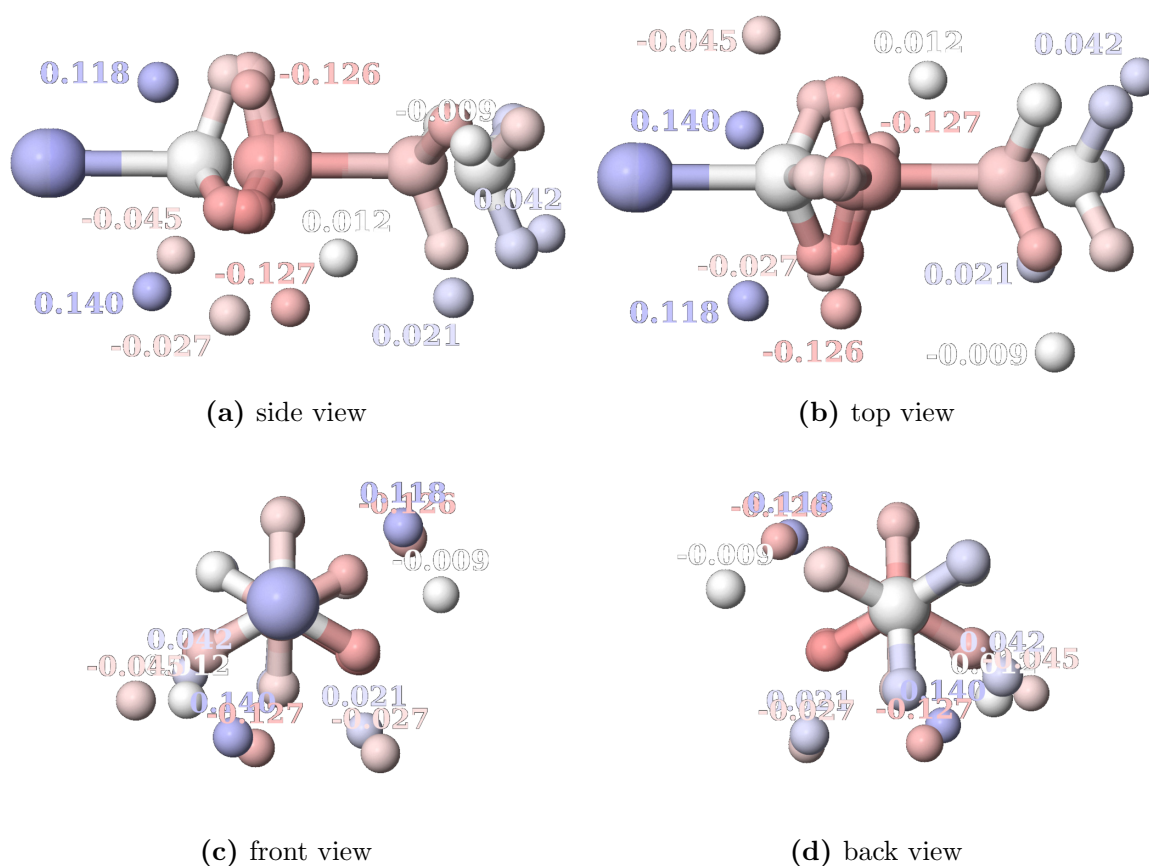


Fig. S51: Single GOCAT individual: four different views of the best rank found during the complete DFT GA runs of cluster c10: r0 for the $N_{\text{Ch}} = 10$ case (2 views already in the main article shown). Values given for the partial charges. Both partial charges and the atoms of the core frames (E, TS, P shown) are colored red/blue in the ranges $[-0.645, +0.645]$ e for charges and $[-43.675, +43.675]$ kcal mol⁻¹e⁻¹ for ESP values, explicitly: Cl: 9.61, 9.94, 9.75; C: 0.97, -10.80, -12.09; N: -0.78, -6.34, -6.47.

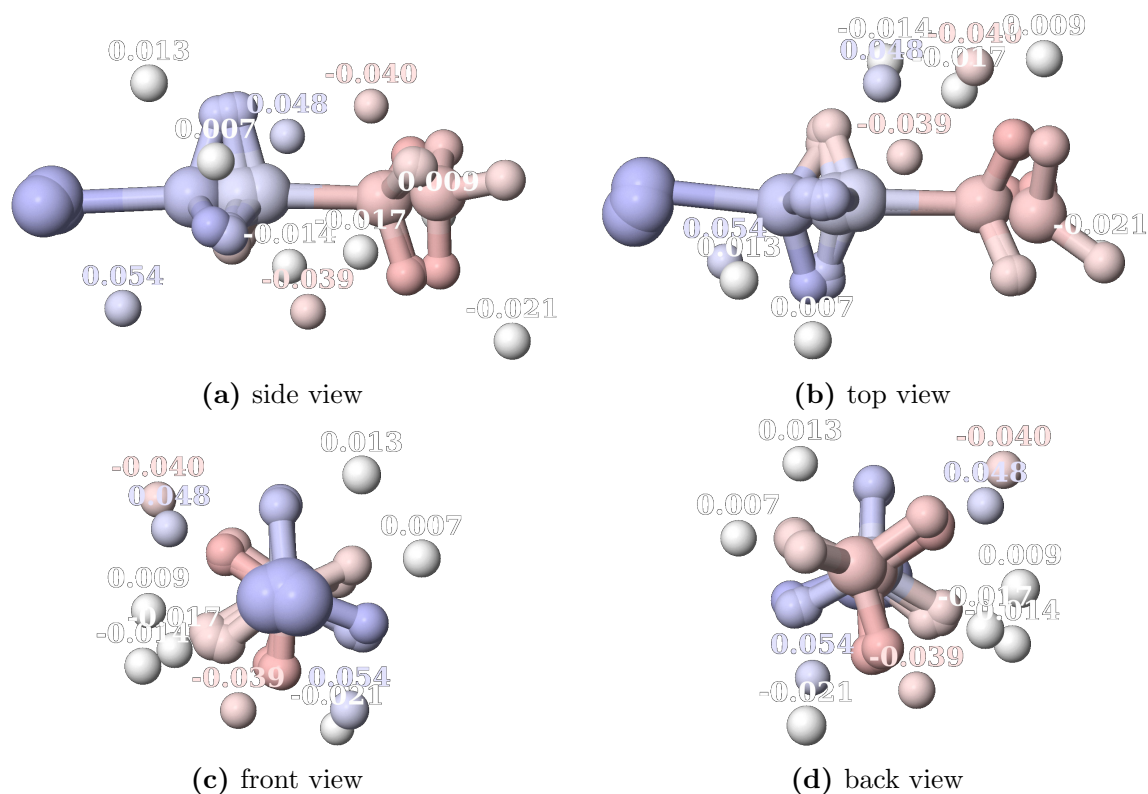


Fig. S52: Single structurally relaxed GOCAT individual: four different views of fully structurally optimized educt, TS and product frames within the electrostatic potential of the GOCAT. Color code follows Fig. S51. Note that a structural relaxation was never applied in all other cases shown here or in the main text. Now the gradient norms are optimized around $1 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ and also the stationary points are checked *via* harmonic analysis (frequencies). Similar to the COSMO PM7 case, the characteristics both of the final ESP values at the core atoms as well as the symmetry breaking pre-oriented educt frame, the “later” TS on the reaction path etc. (main text) can be observed here, too. Main problem, though: As this is a post-GA evaluation, many solutions tend to be “overfitted”, i.e., a full structural relaxation often leads to either loose convergence as no vdW-effects are included or to some “new paths” (new reaction), as for instance the loose degree of freedom (educt) might wander around (and even leave the reaction center). Thus, a structural relaxation of the GOCATs during the GA would be the remedy here.

References

- (S1) Dieterich, J. M.; Hartke, B. OGOLEM: Global Cluster Structure Optimisation for Arbitrary Mixtures of Flexible Molecules. A Multiscaling, Object-Oriented Approach. *Mol. Phys.* **2010**, *108*, 279–291, DOI: [10.1080/00268970903446756](https://doi.org/10.1080/00268970903446756).
- (S2) Dieterich, J. M. OGOLEM.ORG Homepage. <https://www.ogolem.org/>.
- (S3) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 10–20, DOI: [10.1109/MCSE.2007.58](https://doi.org/10.1109/MCSE.2007.58).
- (S4) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (S5) Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **1955**, *77*, 334–338, DOI: [10.1021/ja01607a027](https://doi.org/10.1021/ja01607a027).

Acknowledgements

First and foremost, I would like to thank Prof. Dr. Bernd Hartke for the supervision of the doctoral studies, allowing for utmost creative freedom, the literal open-door-policy, all the trust in my work, and especially for the inception of the **GOCAT** idea from the outset. Contrary to common reactions of most 1st semester chemistry students of today, in fact, I would have already turned my back on chemistry had I not caught a glimpse of something that is called “theoretical chemistry” in my own first semester during Bernd’s undergraduate mathematics lecture. Fortunately, apart from maths, physical chemistry kicked off in the 2nd semester so that the real immersion into chemistry could start to gather pace. Indeed, with a temporal shift of only 1 or 2 semesters in my chemistry studies, this might have taken a completely different direction. I am very grateful that this far-reaching necessary condition of the present work has been met.

I would like to thank Prof. Dr. Carolin König for accepting to be my second referee of this Thesis. Likewise, I would like to thank Prof. Dr. Gernot Friedrichs and Prof. Dr. Ulrich Lünig as further members of my examination committee of the PhD defense.

The financial support of the “Fonds der Chemischen Industrie” (FCI) at the beginning of this project is highly appreciated.

Furthermore, thanks to all my colleagues of the workgroup, to both the current and former members. Especially, I would like to mention Sascha “Sushi” Frick and Christopher Witt for the hard- and software support and hints over all the years.

I am very grateful to Dr. Johannes M. Dieterich for launching the **OGOLEM** framework as a perfect starting point for my own work and for all the discussions over the years.

Thanks to Dominik M. Behrens for further applications within the **GOCAT** theme during his Master’s Thesis, many vivid discussions and for the willingness to carry on the **GOCAT** project to the next level(s) in the future.

I thank Dr. Simone Knief and Dr. Karsten Balzer for the help with the computing facilities at Kiel University.

Many thanks to Prof. Dr. Bernd Hartke, Dr. Henrik R. Larsson, Dominik M. Behrens and Aileen Urban for proofreading the manuscript and further comments.

I am very thankful to Tanja Stojšić and Sonja Ehmke for helping with all the administrative requirements and bureaucracy.

I am very grateful to Dr. Henrik R. Larsson for countless discussions, many free-time activities (art film, orchestra), and actually the whole joint chemistry studies of the last years. He was always a role model for a stringent scientific career and an incentive to try harder. Certainly, your career will culminate as intended!

Last but not least, I am deeply grateful to my significant other, Aileen Urban, without whom this Thesis, the whole studies and life in general would have been utterly impossible! Finally, I thank my family for constituting some type of a welcomed “diametrical opposition” to my scientific life.

Declaration

I, Mark Dittner, hereby declare that the work presented in this Thesis with the title

Globally Optimal Catalysts

was done by me regarding the content and form, under the supervision of Prof. Dr. Bernd Hartke, with no other help than the referenced sources in the text.

This is my first dissertation and the work has never been used in any other dissertation attempts. I have never been deprived of an academic title.

The dissertation complies to the good scientific practice rules as proposed by the German Research Foundation (DFG).

Kiel, September 11, 2019

MARK DITTNER