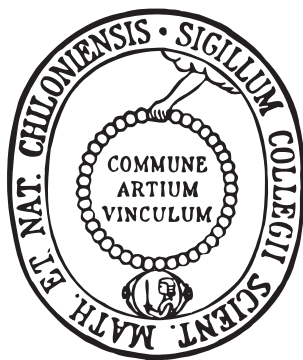

THEORETICAL INVESTIGATIONS OF COVALENT MECHANOCHEMISTRY



Dissertation

in fulfillment of the requirements
for the degree

Dr. rer. nat.

of the Faculty of Mathematics and Natural Sciences
at Kiel University

submitted by

Julian Müller

Kiel, May 2017

First referee: Prof. Dr. Bernd Hartke
Second referee: Prof. Dr. Martin K. Beyer
Date of the oral examination: 6. July 2017
Approved for publication: 6. July 2017

Acknowledgement

First and foremost I express my gratitude to Prof. Dr. Bernd Hartke for the opportunity to work on and finish my doctoral thesis as well as his continued support during the project. His way of letting me free hand in all of my research while being available at any time to discuss difficulties, makes him the best advisor I could wish for.

I am further grateful for the experimental insights and data provided by the collaborating groups of Prof. Dr. Martin Beyer and Prof. Dr. Ulrich Lüning in the scope of the SFB677. Most of my work would not have been possible without the data and feedback by Dr. Doreen Schütze, Dr. Katharina Holz, Dr. Benjamin Lachmann, Dr. Isabel Köhl and Iris Bittner M.Sc.

For their funding of the projects A05 in the SFB677 “Funktion durch Schalten” and Ha2498/12-1, I thank the DFG.

To the remainder of the SFB677 staff: Thank you for fruitful discussions on many poster sessions, summer schools, barbeques, regulars’ table events and other occasions.

I furthermore thank my lecturers for their hard work in ironing out the linguistics bumps in my thesis.

Many thanks to my family, my partner and my friends who always had my back when I needed them.

Last but not least I thank all current and former members of the Hartke group. Thanks for the things I learned, thanks for keeping my sanity and thanks for enduring my sometimes endless ramblings on politics.

Abstract

This thesis is concerned with computational-chemistry investigations of mechanoresponsive molecules which feature predetermined breaking points (PBPs). The mechanophoric systems have been approached at different levels of theory. Reactive molecular dynamics (rMD), density functional theory (DFT), second order *Møller-Plesset* perturbation theory (MP2) and multireference methods were employed to obtain a complete picture of the mechanochemical reactions.

The first of two major subprojects dealt with the mechanochemical behaviour of 1,2,3-triazoles. Within the project A05 of the SFB677 our experimentalist collaboration partners investigated the force-induced, reversed alkyne-azide cycloaddition (AAC) or retro-click reaction. The experiments entailed straining the triazoles embedded in polyethyleneglycole (PEG) chains in single-molecule force spectroscopy (SMFS) experiments in an atomic force microscope (AFM). We used theoretical methods to back up the experimental findings. The structural parameters obtained from DFT calculations and experimental results compared well enough to localize single-molecule covalent events to the very confined region of the molecular PBP. In the subsequent mechanistic investigation of the experiments it was found that the reversed AAC is impossible for the 1,2,3-triazole used and that instead a covalent bond external to the five-ring is broken.

The second project was focused on the application of reactive molecular dynamics to covalent mechanochemistry (CMC). A parameter set for *van Duin's* reactive force field REAXFF was optimized to describe the CMC of disulfide bridges. A reference set has been assembled from MP2 and CASPT2 data for suitable model systems. To our knowledge this is the first and only published REAXFF parameterization based on higher quality data than DFT since the introduction of the formalism more than 15 years ago. It is also the only one that uses multireference referene data which is reliable in the dissociation regions of the PES crucial to CMC. The optimization of the parameter set was done with the evolutionary algorithm (EA) newly implemented in OGOLEM by *Dittner*^[1]. After an extensive trial-and-error phase and with suitable countermeasures to overfitting it was possible to obtain a parameter set with the desired properties. Furthermore a detailed procedure for the global optimization of force field parameters, as well as an extensive set of rational rules specific for the difficult task of fitting many strongly coupled REAXFF parameters simultaneously, was developed and applied. These guidelines are presented in this thesis to guide future REAXFF fitting projects. Subsequently to the optimization the qualities of the parameter set were shown

in proof-of-principle molecular dynamics (MD) simulations for strained mechanophores in vacuo and in solution.

Kurzzusammenfassung

Die vorliegende Arbeit befasst sich mit der Untersuchung von mechanoresponsiven Molekülen, die molekulare Sollbruchstellen (PBPs) enthalten, mit Methoden der Computerverchemie. In Abhängigkeit von den jeweiligen Fragestellungen kamen dabei unterschiedliche Methoden zum Einsatz. Reaktive molekulare Dynamik (rMD), Dichtefunktionaltheorie (DFT), *Møller-Plesset* Störungstheorie zweiter Ordnung (MP2) und Multireferenzverfahren wurden verwendet, um ein vollständiges Bild der mechanochemischen Reaktionen zu erhalten.

Das erste von zwei großen Teilprojekten befasste sich mit dem mechanochemischen Verhalten gespannter 1,2,3-Triazole. Im Projekt A05 des SFB677 wurde von den Experimentatoren die kraftinduzierte umgekehrte Alkin-Azid-Cycloaddition (AAC), auch Retro-Click-Reaktion genannt, untersucht. Hierzu wurden die in Polyethylenglykolketten (PEG) eingebetteten Triazole in Einzelmolekülkraftspektroskopie (SMFS) Experimenten im Rasterkraftmikroskop (AFM) gespannt. Wir verwendeten theoretische Methoden um die experimentellen Resultate zu bestätigen. Die strukturellen Parameter, die mittels DFT-Rechnungen erhalten wurden, erlaubten es, kovalente Ereignisse einzelner Moleküle in der sehr kleinen Region des molekularen PBP zu verorten. In den anschließenden mechanistischen Untersuchungen des Experiments konnte die umgekehrte AAC als Mechanismus ausgeschlossen werden, stattdessen tritt ein kovalenter Bindungsbruch ausserhalb des Fünfrings auf.

Das zweite Projekt befasste sich mit der Anwendbarkeit von reaktiver Molekülmechanik auf kovalente Mechanochemie (CMC). Es wurde ein Parametersatz für die Beschreibung der CMC von Disulfidbrücken mit *van Duins* REAXFF optimiert. Der Referenzdatensatz wurde aus MP2- und CASPT2-Daten geeigneter Modellsysteme zusammengesetzt. Soweit uns bekannt ist, ist dieser REAXFF Parametersatz das erste und einzige publizierte Kraftfeld seit der Einführung des Formalismus' vor über 15 Jahren, der qualitativ höherwertige Referenzdaten als DFT verwendet. Er ist weiterhin der einzige der auf Multireferenzdaten basiert, welche im für CMC wichtigen Dissoziationsbereich verlässlich sind. Nach einer ausgedehnten Trial-and-Error-Phase und mit geeigneten Gegenmaßnahmen gegen Überanpassung war es möglich, einen Parametersatz mit den gewünschten Eigenschaften zu erhalten. Weiterhin wurde eine detaillierte Arbeitsmethodik für die globale Optimierung von Kraftfeldparametern, wie auch ein umfassender Satz von Richtlinien speziell für die schwierige Aufgabe der gleichzeitigen Optimierung vieler stark gekoppelter REAXFF Parameter, entwickelt und genutzt. Dieser

wird in der vorliegenden Arbeit vorgestellt um künftige REAXFF Parametrisierungsprojekte anzuleiten. Im Anschluss an die Optimierung wurden die Qualitäten des Kraftfeldes in beispielhaften Moleküldynamiksimulationen (MD) im Vakuum und in Lösung erprobt.

Contents

1. Introduction	12
2. Theoretical Background	18
2.1. Mechanochemistry and Theoretical Advances	18
2.1.1. Understanding External Forces on Molecules	19
2.1.2. Computational Approaches to Theoretical CMC	26
2.1.3. EFEI - The Isotensional Approach	28
2.2. Reactive Molecular Mechanics	29
2.2.1. Traditional Force Fields	30
2.2.2. Reactive Force Fields	33
2.3. Global Optimization of Parameters Sets	45
2.3.1. Evolutionary Algorithms	48
2.4. Wave Function Methods	54
2.4.1. Basic Approximation and the Hartree-Fock Method	54
2.4.2. Dynamic Correlation and Møller-Plesset Perturbation Theory	56
2.4.3. Density Functional Theory	58
2.4.4. Static Correlation and Multireference Methods	59
2.4.5. The CASPT2 Multireference Perturbation Theory	61
3. Triazole Mechanochemistry	62
3.1. Scope of the Project	62
3.2. Publication: Pinpointing Mechanochemical Bond Rupture by Embedding the Mechanophore into a Macrocycle. ^[2]	64
3.3. Additional Information	69
3.3.1. Elongations of Strained Bicyclic Mechanophores	69
3.3.2. Mechanical Considerations Regarding Mechanically Facilitated Retro Click Reactions	78

4. ReaxFF Parametrization and Disulfide Mechanochemistry	88
4.1. Publication: Efficient Global Optimization of Reactive Force-Field Parameters. ^[1]	91
4.2. Additional Information	104
4.2.1. Automation of Force Field Parameterization	104
4.2.2. Gradients as Additional Property	112
4.3. Publication: ReaxFF Reactive Force Field for Disulfide Mechanochemistry, Fitted to Multireference ab Initio Data. ^[3]	114
4.4. Additional Information	128
4.4.1. Concerning Params Files	128
4.4.2. Strained Molecular Dynamics	141
4.5. Conclusion	143
5. Parameterization of a 1,2,3-triazole force field	144
5.1. Work by Julien Steffen ^[4]	144
5.1.1. The Parameters	144
5.1.2. The Reference Set	145
5.1.3. Results	145
5.2. Further Work	147
6. Summary and Outlook	150
Appendices	161
A. ReaxFF Optimization Input Files	162
A.1. The ffield File	162
A.2. The geo File	165
A.3. The trainset.in File	167
A.4. The params File	169
A.5. The sulfur.ogo File	171
B. Setting up params files	173
C. Triazole Reference Geometries	177

Abbreviations

AAC	Alkyne-Azide Cycloaddition
AFM	Atomic Force Microscope
AS	Active Spane
BBP	Bond Breaking Point
BDE	Bond Dissociation Energy
CASSCF	Complete Active Space Self Consistent Field
CASPT2	Complete Active Space Perturbation Theory Second Order
CI	Configuration Interaction
CMC	Covalent Mechanochemistry
COGEF	Constrained Geometries Simulate External Forces
DFT	Density Functional Theory
EA	Evolutionary Algorithm
EEM	Electronegativity Equilibration Method
EFEI	External Force is Explicitly Included
ESP	Electrostatic Potential
GGA	Generalized Gradient Approximation
HF	Hartree-Fock
HPC	High-Performance Computing
LDA	Local Density Approximation
MD	Molecular Dynamics
MM	Molecular Mechanics
MMC	Metropolis Monte-Carlo
MOES	Multi-Objective Evolutionary Strategy
MR	Multireference
MP2	Second-Order Møller-Plesset Perturbation Theory
PBP	Predetermined molecular Breaking Point
PEG	Polyethyleneglycole
rMD	Reactive Molecular Dynamics
RSPT	Rayleigh-Schrödinger Perturbation Theory
SA	Simulated Annealing
SCF	Self Consistent Field
SMFS	Single-Molecule Force Spectroscopy
SOEA	Single-Objective Evolutionary Algorithm

1. Introduction

With the advent of *Binnig's* and *Quate's* atomic force microscope (AFM) in 1986^[5], a whole new research field became available. Especially the improvements in piconewton instrumentation were responsible for opening a new chapter for covalent mechanochemistry^[6]. It was now possible to manipulate and detect single molecules in the piconewton force regime. Such single-molecular manipulation experiments, as the measurement of the bond strength by *Gaub* and coworkers^[7], became known as single-molecule force-spectroscopy (SMFS) experiments.

The tool used for any SMFS experiment is the AFM. In an AFM (figure 1.1), a substrate is probed using a nanoscopically sharpened needle point, i.e. the tip. The tip is pushed upon a substrate that consequently exerts a force to the tip and the cantilever that holds it. The elastic deformation of the cantilever is measured via a laser beam deflected off its backside. The displacement of the laserbeam on an optical sensorarray is a measure for the deformation of the cantilever, which can be converted to the force acting on the tip by making use of *Hooke's* law.

The setup shown in figure 1.1 differs slightly from the basic setup described above. In contrast to the traditional setup, where the topology of surfaces is sampled horizontally, in SMFS experiments a small number of molecules, ideally a single one, is suspended between the tip and the surface and strained vertically.

For an SMFS experiment the tip starts at its initial position retracted from the surface. The surface is functionalized with mechanophoric molecules. As the tip is lowered into the monolayer of surface-adhered polymeres, their ends can attach to it. The molecule, now suspended between the tip and the surface, is then strained by retracting the tip. There are two possible operating modes for the retraction. The tip is either retracted with a constant speed (force-ramp mode) or positioned to yield a specific straining force (force-clamp mode). While the force-clamp setup mainly allows for conclusions about mechanophore lifetimes, the result of force-ramp experiments are retraction curves that contain information about breaking forces.

Two examples for idealized force extension curves are shown in figure 1.2. The first

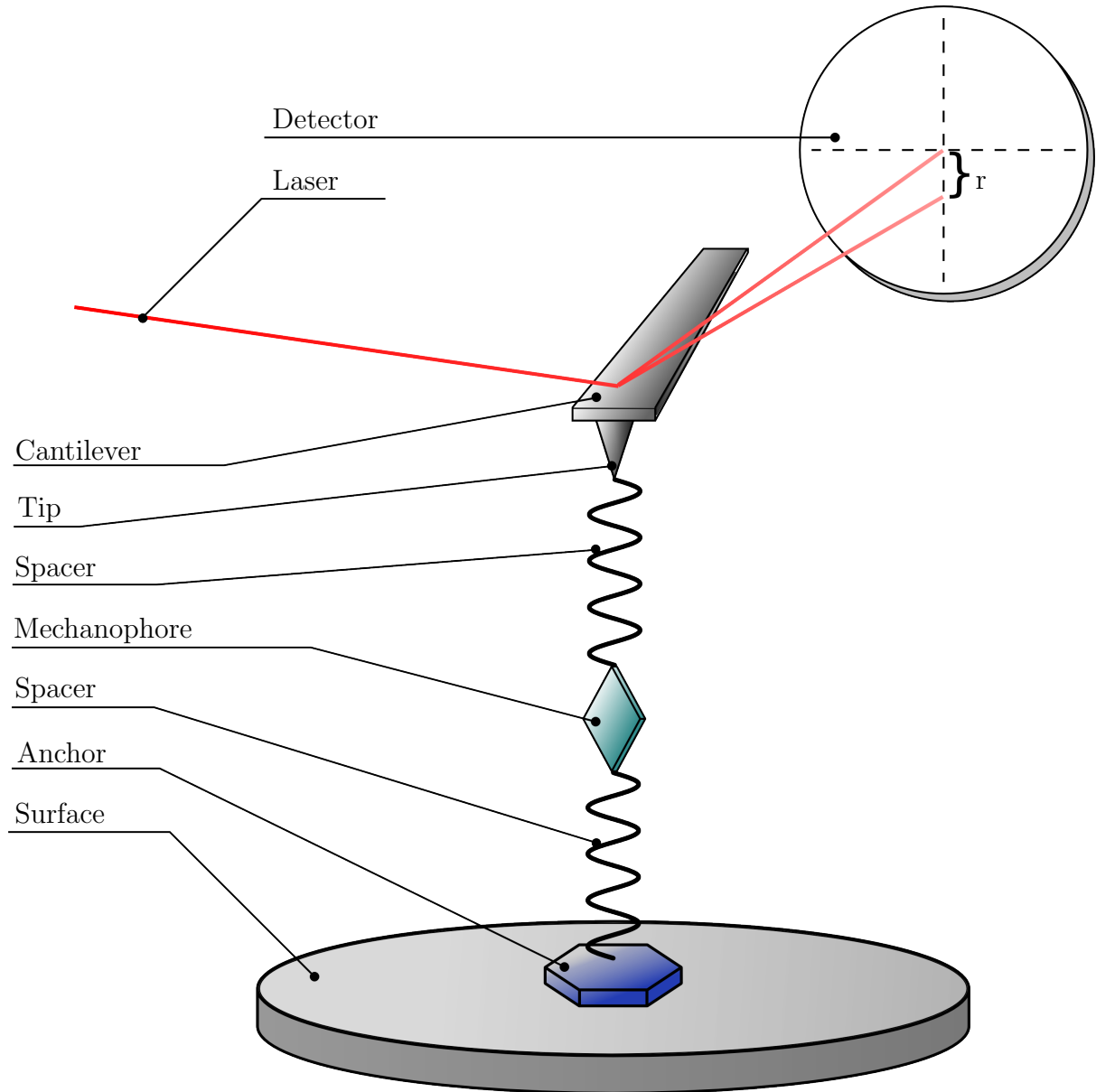


Figure 1.1.: Schematic of an AFM setup for mechanical single-molecular load experiments. The mechanophore is the mechanically active part of the system. It is attached to the surface and the AFM tip using spacer and anchor groups. The cantilever exerts the force on the molecule. A laser reflecting from the back of the cantilever is used to track the path of the tip.

graph in panel a) shows the ideal case of an extension experiment. The tip starts in a position far away from the surface and is approaching it. Since no additional forces are acting on the cantilever during this approaching phase, the measured force is constant. When the tip hits the surface, a force is exerted that bends the cantilever upwards which

1. Introduction

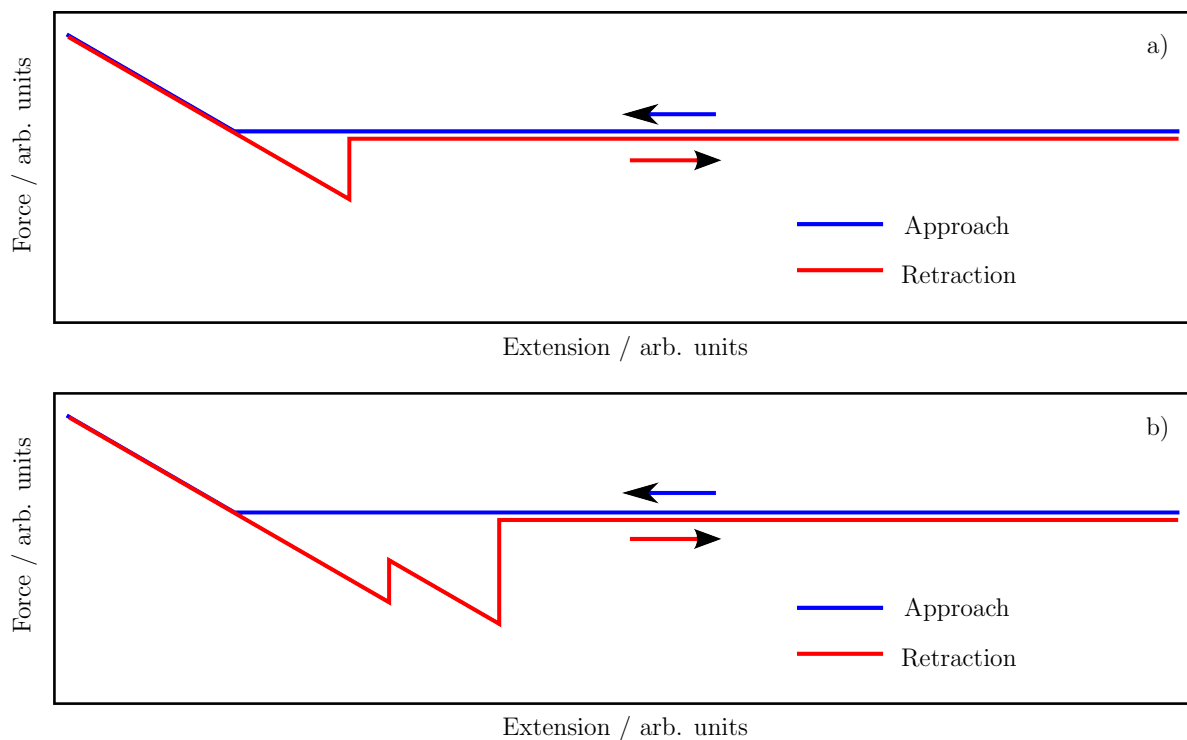


Figure 1.2.: Idealized force extension curves for a single dissociation event in panel a) and a double event in panel b). The curves for the approach of the tip to the surface is drawn in blue while the retraction curves are shown in red. The arrows indicate the direction of the AFM-tip moving relative to the surface.

is the linear increase in the low distance region. When the cantilever is retracted (red curve), the force decreases again. In case the tip did connect with a molecule on the surface, the subsequent stretching of the molecule chain as depicted in figure 1.1 will bend the cantilever downwards, which corresponds to the negative force below the baseline in the force retraction curve. As the molecule becomes more and more extended it eventually reaches a breakpoint where a covalent bond ruptures or a noncovalent rearrangement takes place. When the chain is breaking, the cantilever snaps back to its unbent shape and is now force-free again. The remainder of the retraction curve is just the force-free return to the starting position.

In the SFB677 the SMFS experiment described above was developed one step further. The project A05 of the SFB677 was a close collaboration of groups from the departments of organic chemistry, physical chemistry and theoretical chemistry. Specifically designed macrocycles with bridged mechanophores were synthesized and then experimentally and theoretically investigated. The synthesis was done in the *Lüning* group in the organic

1. Introduction

chemistry department by *Holz, Köhl* and *Bittner*. All AFM experiments were carried out in the *Beyer* group by *Schütze* and *Lachmann*. The computational part of the project was done by myself in the *Hartke* group.

Bridging mechanically active functional groups serves different purposes. For one, it is possible to localize mechanochemical reactions in a single molecule as described below. For another, the bridging or safety line keeps the loose ends after a rupture event in close contact, what may allow for the reversion of the process. Two examples for the macrocyclic systems used are shown in figure 1.3. The mechanical active moiety is the 1,2,3-triazole and the disulfide bridge respectively, these functional groups are bridged by a long alkylic chain which acts as safety line.

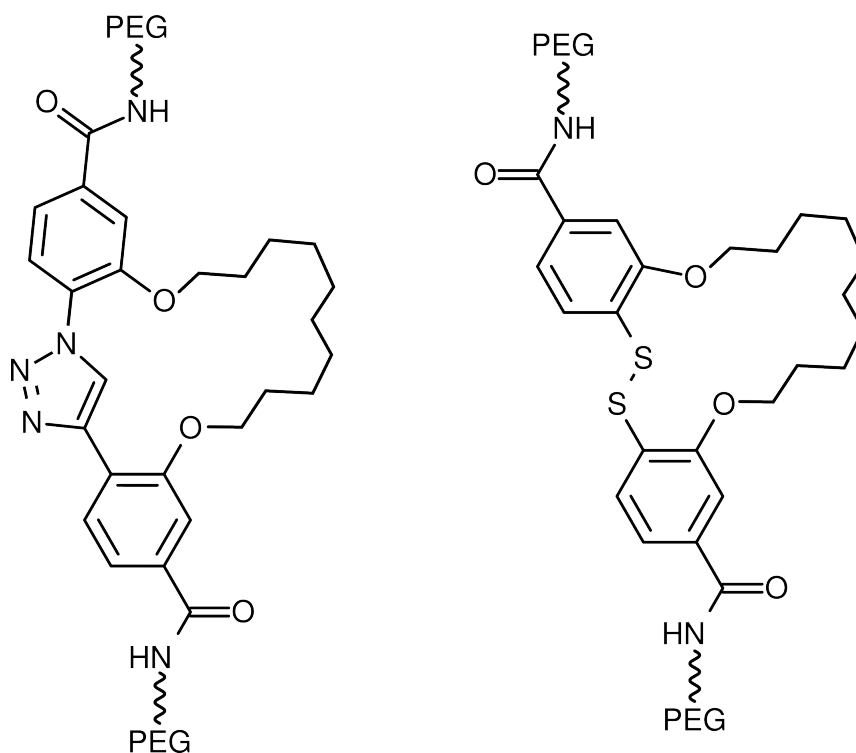


Figure 1.3.: Mechanophoric systems used in this work. The 1,2,3-triazole system on the left was synthesized, experimentally and theoretically investigated for its suspected reversed AAC reaction. Disulfide systems like the one on the right and derivatives were only treated theoretically. They are prototypes for SMJs with switchable molecular conductance.

The second idealized curve in figure 1.2 applies for these systems. Here the situation is shown where the molecular predetermined breaking point (PBP) is bridged by a safety line that sustains the first rupture of the mechanophore. If the bridged mechanophore was addressed in the experiment the connecting molecule between the tip and the surface

1. Introduction

remains unbroken. Following the first rupture event is then the unfolding and straining of the safety line, indicated by the second dent in the force extension curve, up until the second rupture occurs and the tip starts its force-free return to the initial position.

This safety line concept is crucial for the construction of mechanical molecular switches. The ability of the mechanophoric moiety to regenerate after a dissociation event depends on the spatial proximity of the loose ends of the system. This proximity can be ensured by the safety line. The safety line also guarantees that the central mechanophoric moiety was addressed in the experiment, as the second rupture can only occur if the first event was located in the bridged region.

The 1,2,3-triazole-based mechanophoric system was synthesized and used in actual SMFS experiments. Earlier work by *Bielawski* and coworkers indicated that the alkyne-azide cycloaddition (AAC) can be reversed under the influence of a mechanical force^[8]¹. SMFS experiments should be used to observe the force-induced reversed AAC at a single-molecular level. This would have been one of the first chemical reactions that shows the formation of stable nonradical products, after covalent bonds have been broken, on a single-molecular level.

The question posed to the theoretician in this context was if the experimentally observed structural parameters compare well with quantum chemical calculations. The parameter that was experimentally obtained for the reaction is the length of the unfolding safety line in the AFM. If experimental and theoretical elongations compare well, it is a strong indicator that the desired reaction was indeed observed.

As the project progressed it became clear that the reversed AAC is maybe not the best explanation for the experimental results. The covalent single bonds in the bridged region are also strongly strained during the experiments and most likely to dissociate before a reversed AAC occurs. The focus of the theoretical investigations was therefore shifted to mechanistic considerations of strained 1,2,3-triazoles.

The disulfides were investigated for an additional interesting reason. Sulfur atoms have been shown to conduct electrical currents in single-molecular junctions (SMJs)^[9]. Disulfides also have a comparably low bond-breaking force^[10]. The project therefore pursued the question whether it is possible to switch disulfide systems from a conducting to a less conducting or even insulating state with mechanical stimuli^[11]. The safety line concept may then be used to keep the homolytically dissociated sulfur atoms in close spatial contact to each other, which allows for a recombination upon subsiding mechan-

¹The publication by *Bielawski* and coworkers was retracted due to suspected scientific misconduct. This led to a controversy about AFM results on the 1,2,3-triazole obtained within the project A05 later on, which is resolved in this thesis.

1. Introduction

ical stress. Materials that can change their conductivity depending on the mechanical stress may have a wide range of applications in sensor technology or future computing devices.

To investigate the reactivity and dynamics of a strained disulfide moiety *van Duin's* reactive empirical potential REAXFF^[12] was refitted. With this force field it will be possible to simulate the behaviour of disulfide-based molecules in SMFS experiments or bulk materials under external strain on significantly longer time scales than with ab-initio MD oder direct DFT dynamics.

In addition to this application the perspective is intriguing from the theoretical point of view. The theoretical treatment of CMC reaction remains a challenge today^[13], since the experiments take place at time scales of seconds and in solution at finite temperatures and involve breaking and formation of covalent bonds. Typically dynamics of large systems, i.e. systems containing more than 1000 atoms, over long simulation times are treated with force fields which usually do not allow bond formation. Simulations of this size are untractable with modern ab-initio or DFT methods, which could deal with the bond formation. This dilemma can be solved by fitting and using empirical reactive potentials like REAXFF for the description of CMC. In this work, I therefore pursued the refitting of REAXFF parameters to reliable reference data. This made the tools available which are necessary to simulate SMFS experiments at all relevant time- and system scales.

2. Theoretical Background

2.1. Mechanochemistry and Theoretical Advances

It is rather difficult to pinpoint the origin of mechanochemistry in the history of chemistry. Fundamental, although empirical, concepts were already known some 2000 years ago, but real effort in mechanochemistry as a field of its own started only in the second half of the 19th century^[14]. The term mechanochemistry itself goes back to *Wilhelm Ostwald* who coined it in the early 20th century in his textbook on general chemistry^[15]. There he mentioned it as a field of chemistry besides the then more commonly known disciplines of thermochemistry, electrochemistry and photochemistry.

It is pretty self-explanatory that mechanochemistry entails all chemical reactions that are induced with a mechanical force as a stimulus. In times mechanochemistry was used in different fields of research without knowing of one another, and maybe even without noticing that mechanochemical concepts were used. Today mechanochemistry is not a unified field of research with a general methodology but is instead practiced throughout all disciplines of chemistry^[14].

Ball milling techniques are utilized in material sciences and inorganic chemistry^[16]. Sonochemistry is used for synthesis in organic and inorganic chemistry and very closely related to mechanochemistry^[14,17]. Stress responsive materials react to mechanical strain and can be used for sensors or intelligent materials^[18]. Many more applications can be thought of. For more detailed information on the field of mechanochemistry the reader is referred to the reviews by *Laszlo Takacs*, *Jordi Ribas-Ariño* and *Dominik Marx*.^[13,14]

Part of the unification problem mentioned above is that the underlying principles of mechanochemistry are not fully understood yet, neither from the empirical nor the theoretical point of view. A new experimental approach to shed light on the fundamental mechanisms of mechanochemistry was introduced by *Hermann Gaub* and coworkers. Their influential paper titled “How strong is a covalent bond”^[7] opened a new branch of research. An SMFS method was proposed in which single molecules suspended in an AFM are probed under a defined mechanical strain. The force necessary to break a

2. Theoretical Background

single covalent bond was measured and validated against a theoretical model.

The theoretical foundations of this approach to covalent Mechanochemistry (CMC) will be discussed in the following section. Due to its significance as a model for covalent bonds, the findings in the following section will mainly revolve around analytical investigations of the *Morse* potential. However the principles are generally applicable and can be used with any potential energy function available.

2.1.1. Understanding External Forces on Molecules

Qualitative understanding of CMC requires rigorous models for the phenomenological description of force-dependent single molecular processes.

Very early work on such models was done by *Eyring* and coworkers in the form of a force-dependent extension to their transition state theory^[19]. More than two decades later *Zhurkov* proposed a model for strained solids based on *Arrhenius*' equation^[20]. Although the work was published well before the year 1978, models of this type became widely known as *Bell* type models. In his work *George Bell* used an almost identical equation as *Zhurkov* to quantify cell adhesion forces^[21].

Bell's model assumes that the activation energy for a reaction is reduced by a linear term dependent on the force acting along the reaction coordinate. The problem with this simplistic model is that it only accounts for a linear force-dependent shift of the equilibrium position along the reaction coordinates and neglects any distortion happening on the PES. Furthermore, any reaction treated by the model must be able to be projected to a one dimensional reaction coordinate. The limitations arising from these weaknesses are discussed in many publications^[10,13,22,23].

Since *Bell's* publication in 1978 and especially since *Gaub's* paper in 1999, a manifold of models for CMC and mechanochemistry in general was proposed to address the issues mentioned^[7,10,22,24-27]. All of them can be viewed as more or less sophisticated extensions to *Bell's* model.

The model presented in this section will combine a tilted potential approach with the principles of *Bell's* model. This approach provides an intuitive understanding of the effect of external mechanical forces on potential energy surfaces and the subsequent molecular dynamics. The approach was in slight variations already used in different research groups^[7,22,24,25]. The intricacies and distinctions between all the different varieties of the models will be omitted from here on to provide a comprehensible introduction to mechanochemistry, rather than an obfuscating review of all details to CMC. The model is also sufficiently accurate to explain the findings discussed in later chapters of this

2. Theoretical Background

thesis. Readers who are interested in details of the models are referred to *Marx*' review and references cited therein^[13].

The tilted Morse Potential

The one-dimensional *Morse* potential is a standard model for the dissociation energy curve of a covalent chemical bond. The potential in equation 2.1 has three parameters. The dissociation energy D_e , the force parameter β and the equilibrium distance x_0 . In the example considered here D_e is 250 kJ/mol, β equals 2 \AA^{-1} and x_0 is set to 2 \AA . As it becomes important later on, it should be noted that the reduced mass of the oscillator is assumed to be 16 a.m.u. and therefore the vibrational frequency is $5.7 \cdot 10^{13} \text{ s}^{-1}$.

$$V(x, F) = D_e(1 - e^{-\beta(x-x_0)})^2 - F \cdot (x - x_0) \quad (2.1)$$

The potential above already includes a distortion of the potential, i.e. it is tilted, by an external force F acting along the coordinate x . The distortion term follows directly from the relationship between the potential energy and the force acting on a particle, given by the negative gradient of the energy. The negative first derivative of the tilted *Morse* potential (eq. 2.1) yields the additional constant force F which governs the mechanical strain.

The impact on the shape of the potential energy curve is profound. In figure 2.1 three different situations are shown. The force-free potential curve ($F = 0$) has the typical shape that is familiar to the chemist's eye. From the force-free potential a critical force F_c may be derived which is the absolute value of the maximum of the first derivative. This critical force F_c is the force necessary to break the bond. The critical force in a *Morse* potential is $F_c = \beta D_e/2$.

When applying a positive force, i.e. a force directed to stretch the bond, the shape of the potential curve changes. A local maximum occurs along the bond coordinate and the equilibrium distance is shifted towards higher values of x . There are two distinct implications to this. The equilibrium geometry of molecules is distorted when subjected to mechanical strain and beyond the emerging maximum, the potential becomes repulsive. This maximum is called bond breaking point (BBP) and qualifies as a first order saddle point or transition state. There are two quantities associated with the BBP: The bond breaking length which is the internuclear separation that is needed to split the bond apart and the height of the barrier which is smaller than the dissociation energy D_e of the potential. The barrier vanishes completely when the applied force equals the

2. Theoretical Background

critical force $F = F_c$.

The last case shown in figure 2.1 is a negative force directed to shrink the bond. Unsurprisingly the equilibrium distance is shifted to smaller internuclear separations. Furthermore the potential becomes attractive in the asymptote as the applied force will always drive the nuclei back together.

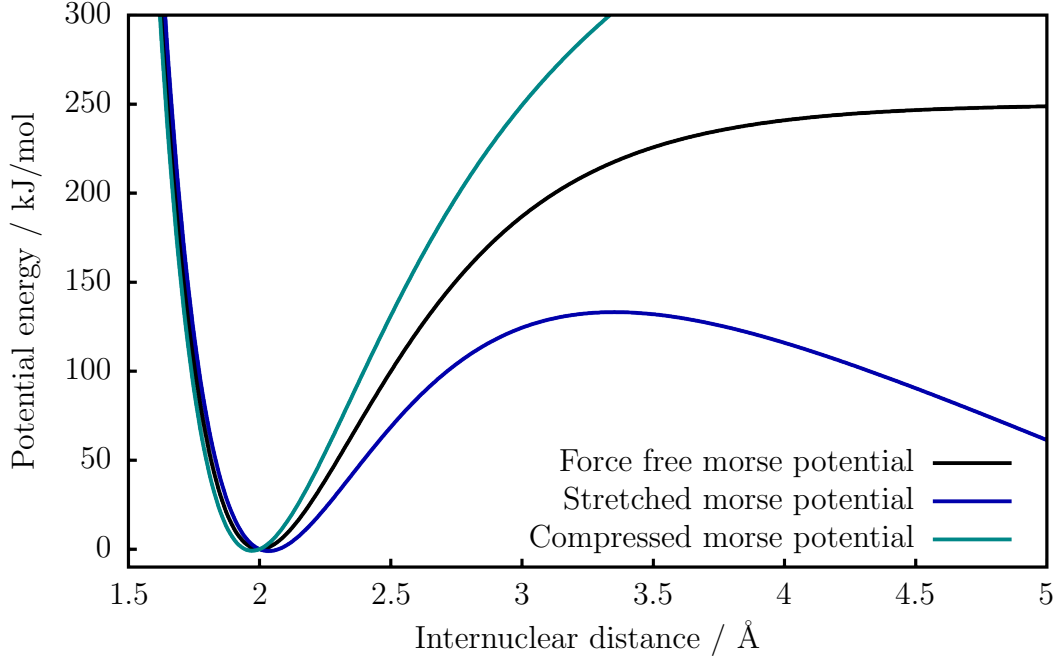


Figure 2.1.: Various *Morse* potentials with additional external force terms applied. Force-free situation (black), a *Morse* potential with a positive force of $0.25F_c$ applied (blue) and a *Morse* potential with a negative force load of $0.25F_c$ (cyan).

It has already been mentioned that the height of the potential energy barrier, which emerges when force is applied, defines the force-dependent activation energy for the dissociation. Since the equilibrium position x_{eq} and the bond breaking length x_{bp} are shifted depending on the applied force, both need to be calculated to find the analytical expression for the barrier height (eq. 2.3). Equation 2.2 computes the force dependent roots of the negative first derivative of the distorted *Morse* potential (eq. 2.1), which correspond to the quantities x_{eq} and x_{bp} .

$$x(F) = \frac{1}{\beta} \ln \left(\frac{\beta D_e}{F} \pm \sqrt{\frac{\beta D_e}{F} \left(\frac{\beta D_e}{F} - 2 \right)} \right) \quad (2.2)$$

As already pointed out by *Uggerud* and coworkers^[22] this equation has multiple

2. Theoretical Background

regimes with a varying number of valid solutions in x .

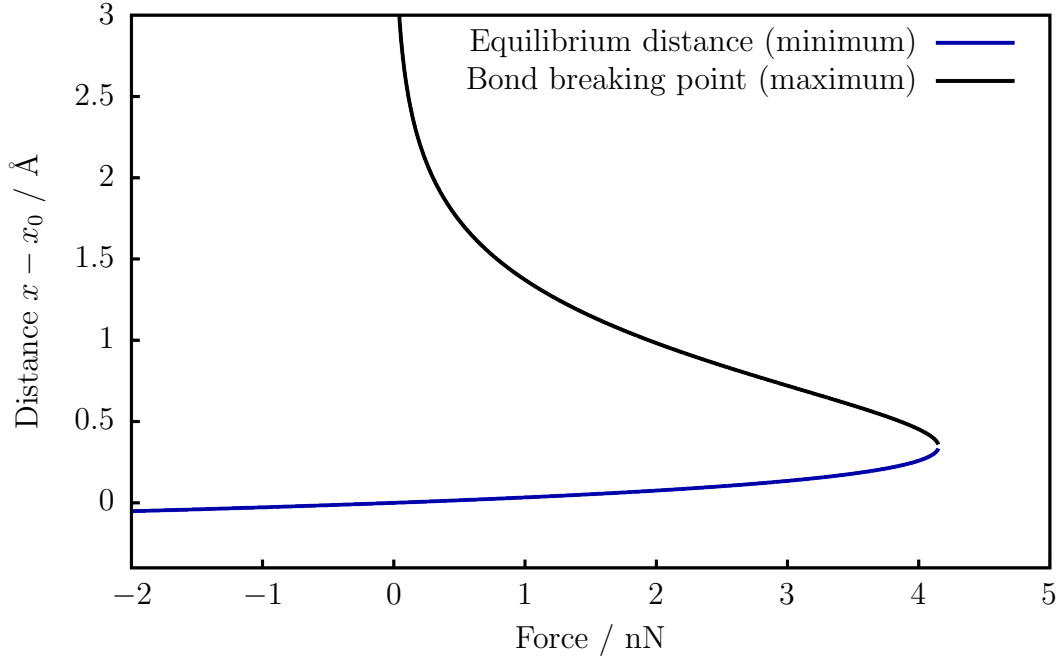


Figure 2.2.: Roots of the negative first derivative of the *Morse* potential with respect to $x - x_0$. The lower blue line corresponds to the shift of the equilibrium distance x_{eq} with respect to x_0 , the relative position of the breaking point x_{bp} is plotted in black.

In the region in figure 2.2 for $F \leq 0$ there is only one solution. This solution is the relative shift of the equilibrium distance x_{eq} with respect to the force free equilibrium x_0 . In the second region for $0 < F < F_c$ there are two solutions. The shift of the equilibrium distance (blue) and the position of the bond breaking length x_{bp} with respect to x_0 . Beyond the critical force, $V(x, F)$ is repulsive for all x and equation 2.2 has no real solutions.

The effective barrier height $\Delta V_{eff}(F)$ is the potential energy difference for $V(x_{bp}, F)$ and $V(x_{eq}, F)$. Thus it is only possible to calculate the barrier height $\Delta V_{eff}(F)$ in the second regime where $0 < F < F_c$. Inserting the expressions for x_{bp} and x_{eq} obtained in equation 2.2 into the potential function and calculating the difference $V(x_{bp}, F) - V(x_{eq}, F)$, yields the expression for the effective barrier height in equation 2.3.

2. Theoretical Background

$$\begin{aligned}\Delta V_{eff}(F) &= V(x_{bp}(F)) - V(x_{eq}(F)) \\ &= \frac{F}{\beta} \left(\sqrt{\frac{\beta D_e}{F} \left(\frac{\beta D_e}{F} - 2 \right)} - \ln \left(\frac{\beta D_e}{F} + \sqrt{\frac{\beta D_e}{F} \left(\frac{\beta D_e}{F} - 2 \right)} - 1 \right) \right) \quad (2.3)\end{aligned}$$

The form of this barrier decay depends on the potential energy function. In case of the *Morse* potential the result is the rather complex exponential decay shown in figure 2.3.

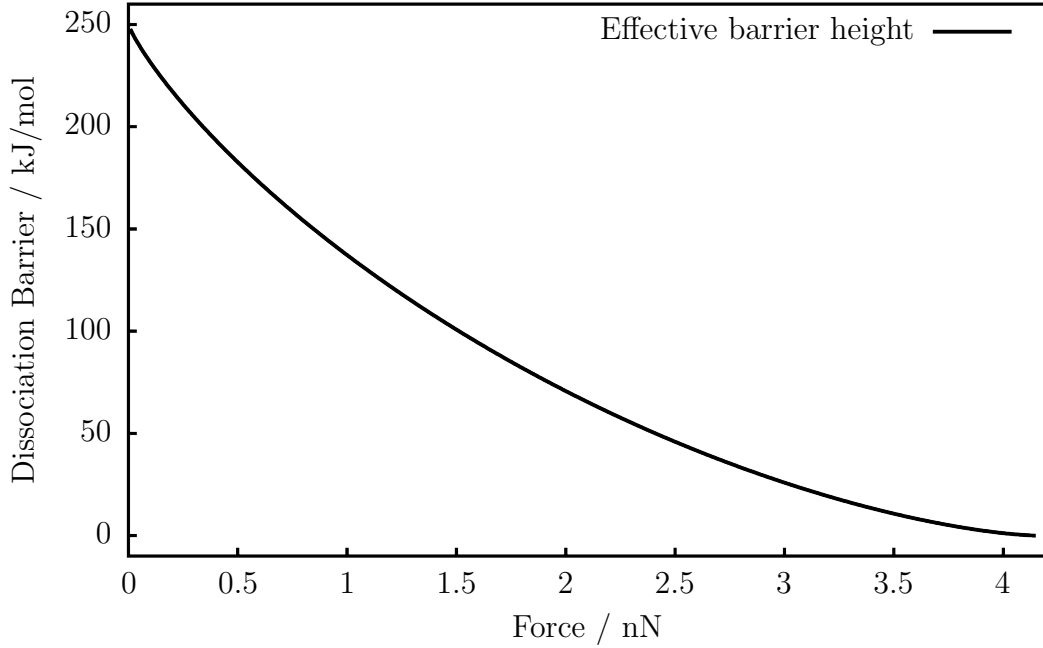


Figure 2.3.: Decay of the effective barrier for the dissociation of a covalent bond from $V_{eff}(F) = D_e$ at $F = 0$ to $V_{eff}(F) = 0$ at $F = F_c$.

With an expression for the activation barrier at hand it is possible to obtain estimates for the force-dependent dissociation rates by plugging the activation energy in *Arrhenius'* equation.

$$k(F, T) = k_0(T) \exp \left(-\frac{V_{eff}(F)}{RT} \right) \quad (2.4)$$

The rate constant $k(F, T)$ is now explicitly force- and temperature-dependent. The rate $k_0(T)$ quantifies the force-free dissociation rate of a bond. There have been discussions to some extent as to how to calculate the force free-rate constant $k_0(T)$. Initially it

2. Theoretical Background

was assumed that $k_0(T)$ is the vibrational frequency of the strained bond^[13,21] but *Evans* has shown that this preexponential factor has to be several orders of magnitude smaller than the oscillation frequency^[28]. In a recent publication by *Uggerud* and coworkers, they found TST estimates of $k_0(T)$ to be unphysically high^[22]. It can be concluded that an estimation of the force-free rate constant is not a trivial task. However, for a qualitative consideration of the CMC processes the preexponential factor is not relevant and will be regarded as being unity. With $k_0 = 1$ equation 2.4 is confined to a range of $0 \leq k(F, T) \leq 1$.

Plotting the rate constant $k(F, T)$ over F for different temperatures as it was done in figure 2.4 reveals the sigmoidal shape of the reaction rate. The shape of the curves imply that neither experimental results nor activation forces obtained from MD simulations at finite temperatures can be expected to be infinitely sharp values. Furthermore the forces at which reaction can be observed may be considerably lower than the critical force F_c . The critical force is merely the upper limit for the force beyond which every trajectory dissociates instantaneous.

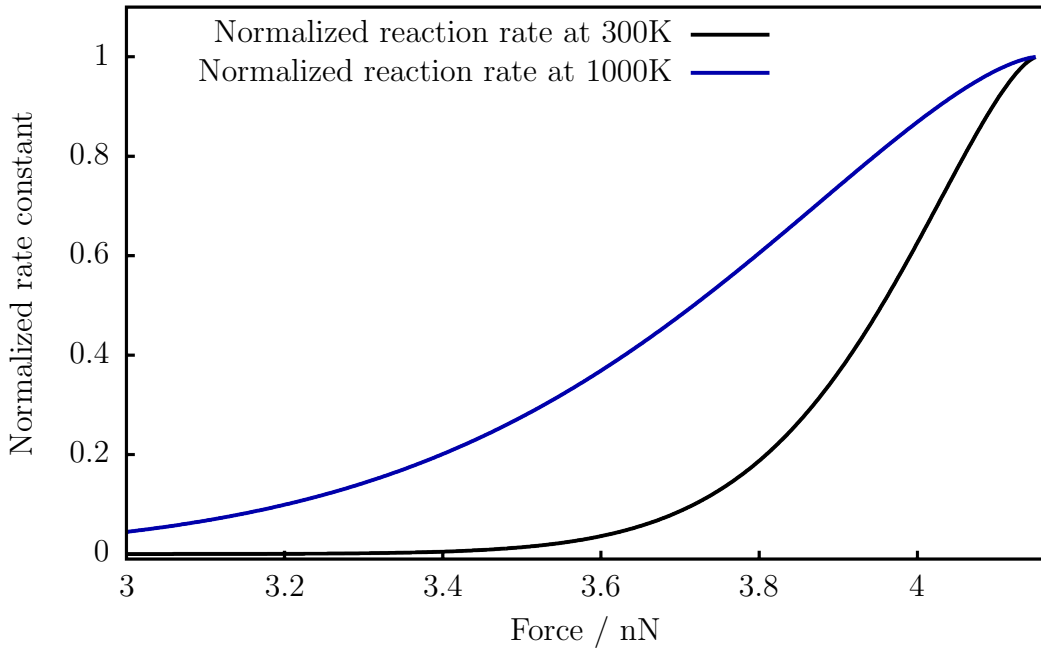


Figure 2.4.: Normalized reaction rates (units of k_0) (reaction probabilities for trajectory ensembles) for the covalent bond dissociation in the *Morse* potential. At rates $k \geq 0.5$ the reaction becomes more probable than the chance of the system to remain in the educt state.

The idea of the activation force may be refined further to coincide better with values

2. Theoretical Background

found in experiments or molecular dynamics simulations. This is desirable since the critical force F_c for a reaction is just the upper limit for the experimentally obtained breaking force. Usually experimental breaking forces are just a fraction of the critical force due to dynamic effects. The activation force is chosen to be the force for which $k(F, T)$ becomes $k_0/2$. This is because at that force the probability of a system to react within the lifetime $\tau_0 = 1/k_0$ is greater than 0.5, which means the system is more likely to react than not.

For the *Morse* potential considered here (eq. 2.1), the activation force would be 3.95 nN at 300 K and 3.72 nN at 1000 K.

When bond-breaking forces are computed as described here, there will still remain a large discrepancy between theoretical predictions and experimental results. This is because the preexponential factor, lifetimes and experimental timescales have not been taken into account yet. As pointed out by *Marx*, the force-loading rates in SMFS experiments are small on the timescale of molecular vibrations. A typical force-loading rate of 10 – 100 nN/s means that a molecule with a bond breaking force of 1 – 5 nN suspended in the AFM is strained for 0.01 – 0.5 s before a dissociation event occurs. This is twelve orders of magnitude slower than molecular vibrations on average. Although *Evans* argues that the force-free rate constant is at least thousandfold smaller than the molecular vibrations^[13,28], there remains a gap of at least eight orders of magnitude between experimental timescales and the dissociation rates.

Figure 2.5 shows the relation between the applied force and the lifetime of the strained bond. Opposed to the situation where k_0 was assumed to be 1, for the lifetimes an estimate for k_0 is needed that at least reflect the physical reality to some degree. Therefore the lifetimes τ were calculated as $\tau = 1/k(F, T)$ where, according to *Evans* findings, $k_0(T)$ was assumed to be one thousandth of the vibrational frequency of the Morse potential. The *Morse* parameters from section 2.1.1 yield $k_0 = 5.7 \cdot 10^{10} \text{ s}^{-1}$ and $\tau_0 = 1.8 \cdot 10^{-11} \text{ s}$. These values were chosen to accommodate for the quantity of the preexponential factor discussed above. Lifetime plots of this kind have been used before by *Beyer*^[24].

The activation force decays rapidly with the lifetime. On the timescale of milliseconds and at room temperature the activation force is 2.59 nN, which is significantly lower than the estimated 3.95 nN above or even the critical force of 4.15 nN. Furthermore the activation forces show a very strong temperature-dependence on the experimental time scales of milliseconds to seconds. For a representative picture of CMC, computational methods are required to simulate these time scales.

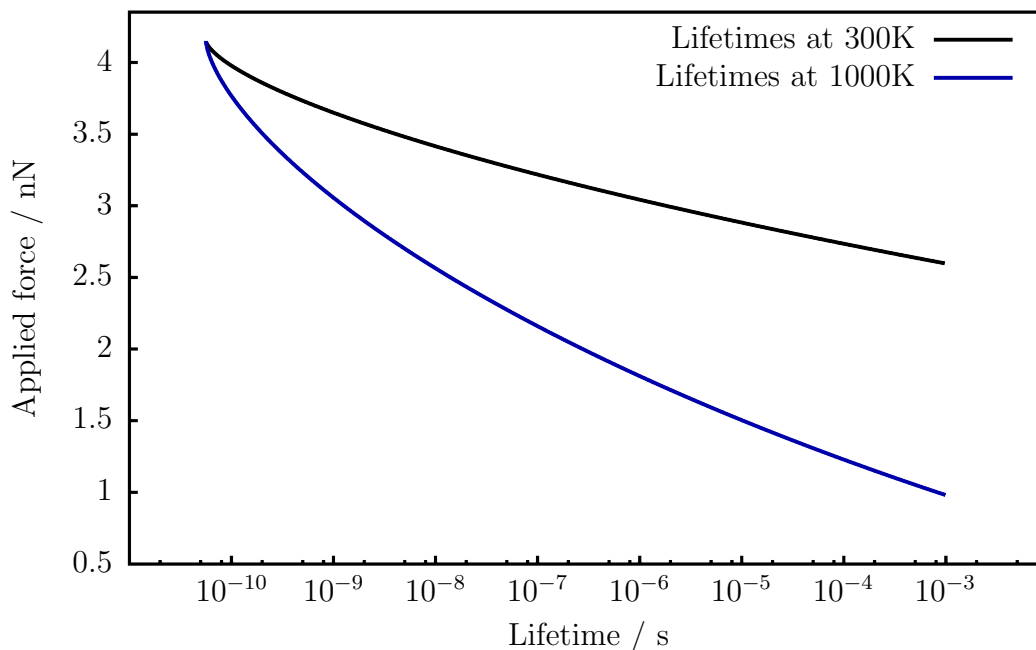


Figure 2.5.: Activation force plotted over the lifetime of a strained bond. The lifetimes τ are calculated as $\tau = 1/k$ and k_0 is approximated as one thousands of the vibrational frequency of the morse oscillator.

Further effects can influence the reaction rates significantly but are not considered in this simplistic model, as it has not been necessary to compute properties to a higher accuracy in this thesis.

2.1.2. Computational Approaches to Theoretical CMC

The discussion above was dedicated to impart the basic concepts of mechanochemistry. In the process it was implied that the external force F is already acting on the molecule in one way or another. Up to now no efforts have been made to explain how external forces are dealt with in computational models.

External forces are usually included into computational models using one of the following two approaches. The isometric approach, which includes forces indirectly by using geometric constraints and the isotensional approach, which includes the force explicitly as an additional term to gradients or potential energies.

COGEF - The Isometric Approach

The acronym **COGEF** stands for **CO**nstrained **G**eometries simulate **E**xternal **F**orces and was introduced by *Martin Beyer* in 2000^[24].

The method is a protocol based on a relaxed potential energy surface scan of the strained molecule. An internal coordinate, typically an interatomic distance, is fixed and set to a certain value r_0 , while all other degrees of freedom in the system are allowed to relax locally. The COGEF potential is obtained by a stepwise increase of r_0 and a subsequent geometric relaxation at every point.

Since all degrees of freedom are locally optimized at any point r_0 the only remaining non-zero contribution to the gradient of the PES is the force acting along the fixed, i.e. the strained, coordinate. The mechanical force acting on the molecule to induce a certain distortion can then be extracted as the length of the gradient vector, or as the first derivative of the COGEF potential.

Further information that can be obtained from the COGEF potential are the critical force F_c and the bond-breaking distance that were discussed before. The critical force is the maximum of the first derivative or the point of inflection of the COGEF potential. The bond breaking distance is the corresponding coordinate value.

COGEF is a straightforward intuitive approach to extract the mechanochemical information mentioned from any theoretical method that provides gradients and energies. From the practical point of view the method can be easily applied, as most quantum chemical or molecular mechanics software packages have constrained geometry optimizations implemented. The potential itself may then be used for a phenomenological analysis as the one described in section 2.1.1.

However the isometric approach has some shortcomings that can be resolved by isentensional methods. First, the external force is not chosen explicitly but follows indirectly from the gradient of the potential. If molecular distortions at certain fixed forces are needed, more than just one calculation to find the desired force may be necessary. Second, geometric constraints are a rather unsuitable method to introduce force in molecular dynamics simulations except for very special applications. Even though the COGEF formalism wasn't intended to be used for MD applications it's still noteworthy that this approach can not be used to conduct the force strained reactive molecular dynamics simulations that were a major part of this work.

2.1.3. EFEI - The Isotensional Approach

The **EFEI** method (**E**xternal **F**orce is **E**xplicitly **I**ncluded) was introduced by *Marx* and coworkers in 2009^[29]. The approach operates a little more in the spirit of the model put forward in section 2.1.1 as it incorporates the force directly as an additional term to the gradient vector of the PES. The potential is therefore tilted in the correct fashion and remains at its full dimensionality without projecting the CMC reaction to a single reactive coordinate.

First it should be recalled that the force on any particle moving in a potential is given by the negative gradient vector $\vec{F} = -\vec{\nabla}V(\vec{r})$ of that potential. An additional force to an arbitrary number of atoms (usually two anchoring atoms) can now be introduced simply by adding a force vector \vec{F}_0 to the gradient.

Similar to the COGEF approach, EFEI in principle works with any method that computes potential energies and gradients. The challenge with this method is a very practical one. In most quantum chemical or MD packages the user has no immediate access to the gradient while a calculation is running. This inability to manipulate gradients during optimization steps makes it hard to realize EFEI calculations. An exception to this is the TURBOMOLE^[30] package that allows access and modification of the molecular gradients before they are used by the software. For molecular dynamics simulations additional force vectors can be added in the LAMMPS^[31] suite to an arbitrary number of atoms. Many other software packages would need some coding effort to include explicit forces in the gradients or Hamiltonians respectively.

When the technical obstacles are dealt with, it is in contrast to the COGEF formalism very easy to obtain the molecular geometric distortion for a specific force. The procedure consists of a standard geometry optimization with the external force vector for the two (or more) atoms being added to the gradient of the potential.

It is furthermore the intuitive approach to use for MD simulations of SMFS experiments conducted on mechanophores. The additional force vector acting on the anchoring atoms will be added in each propagation step and the trajectories evolve on the tilted potential energy surface.

As pointed out by *Marx*, the EFEI- and the COGEF potential are *Legendre* transforms of one another, with r_0 and F_0 being conjugate variables^[13,29]. Therefore both approaches will yield the same results for similar problems. This means a COGEF calculation for a certain r_0 returns a straining force F , using the same force F for an EFEI type optimization will yield the distance r_0 for the strained coordinate again. It was already mentioned that the computational demands with both approaches vary depending on

the problem. For example just one calculation with EFEI is needed to get a molecular geometry at a certain straining force while the same information would require at least a few points along the COGEF potential. Because of the consistency of the results and the different computational demands both methods complement each other very well in investigating CMC.

2.2. Reactive Molecular Mechanics

The aim of molecular dynamic (MD) simulations is to solve *Newton's* equations of motion for a set of nuclei moving on a potential energy surface. To accomplish this task, simple propagation algorithms like the *Verlet* integration (eq. 2.5) may be used^[32].

$$\vec{r}_i(t + \Delta t) = 2\vec{r}_i(t) - \vec{r}_i(t - \Delta t) - \frac{\vec{\nabla}V(\vec{r}_i)}{m_i}(\Delta t)^2 \quad (2.5)$$

Verlet's equation calculates the new position of the i th nucleus \vec{r}_i at the time $t + \Delta t$ from its current position at time t , its position at time $t - \Delta t$ and the force $\vec{F}_i = -\vec{\nabla}V(\vec{r}_i)$ acting on it. Propagating the geometry stepwise through time yields the trajectory in the $6N$ -dimensional phase space, where N is the number of nuclei.

The acting force that is the gradient of the system has to be calculated for every subsequent timestep. These timesteps have to be small in order to capture the fastest movements in in molecular system. For most molecular systems the fastest processes are hydrogen vibrations. To simulate them, timesteps Δt of 0.1-1.0 fs are necessary. As modern MD simulations are routinely used to capture effects on the nanosecond timescale, single trajectories usually consist of 10^6 - 10^7 timesteps.

Consequentially the high amount of timesteps and therefore computations of the gradient vector is what makes MD calculation a time consuming task. In fact MD simulations are computationally so demanding that ab-initio quantum chemical methods are not eligible for them. Ab-initio and DFT approaches have time scalings between $O(M^3)$ and $O(M^7)$ depending on the method, where M is the number of basis functions. With increasing system size they become unfeasible rapidly. This means simulations on the nanoseconds timescale can only be done for the smallest systems with small basis sets using DFT.

To set this in perspective consider state of the art calculations recently published by *Martinez* and coworkers^[33]. There, a nanoreactor containing 228 atoms was propagated for 1296 ps using an approximate *Hartree-Fock* ansatz. The resources needed for that

2. Theoretical Background

massive parallel simulation amounted to 132,400 CPU/GPU hours or just above 15 years of computation time.

Advancing into the region of multi-nanosecond or even milisecond simulations for system containing several thousand or millions of atoms requires a much more efficient approach. Here molecular mechanics methods come into play.

In molecular mechanics the quantum mechanical nature of chemistry is neglected and the potential energy surface is approximated by analytic functions of varying complexity with analytic gradients. Nuclei are assumed to be classical particles evolving on these PESs according to *Newton's* laws of motion. Depending on the size of the system the largest number of contributions is the pairwise nonbonding interaction between all atoms. The formal scaling of MM methods is therefore $O(N^2)$. Cutoff distances and other simplifications like the *Ewald* summation or the improved particle mesh *Ewald* PME method can lower this quadratic scaling to $O(N \ln N)$. Furthermore, the prefactor O is several orders of magnitude smaller than in the case of ab initio simulations.

2.2.1. Traditional Force Fields

The discussion of force field methods will be started off by highlighting the foundations of molecular mechanics and traditional force fields. The findings will then be used to discuss similarities and differences for reactive formalisms.

Traditional force fields are built on the notion that the total energy of a molecular system can be decomposed into energetic contributions from one-body-, two-body- up to n-body terms. All these contributions are governed by the parameterized functions summarized in equation 2.6.

$$E_{\text{tot}} = E_{\text{bond}} + E_{\text{val}} + E_{\text{tors}} + E_{\text{coul}} + E_{\text{vdw}} + E_{\text{cross}} \quad (2.6)$$

The total energy E_{tot} is composed of the bonding energy E_{bond} , the angle contributions E_{val} , torsional term E_{tors} and the nonbonding *Coulomb* interactions E_{coul} and *Van-der-Waals* interactions E_{vdw} . The cross terms E_{cross} were added to form the so called class II force fields and describe couplings between the other five energy terms. The cross terms will not be discussed in detail here but further information can be found in the literature^[34,35].

2. Theoretical Background

Bonded Energy Terms

The most simple and common form of a bonding potential is a harmonic approximation to the bonding energy. The bond energy contribution for every bonded pair of atoms is computed according to equation 2.7.

$$E_{\text{bond}} = \sum_i \frac{1}{2} k_i (r_i - r_{i,0})^2 \quad (2.7)$$

Here the force constant k_i and the equilibrium distance $r_{i,0}$ are the empirical parameters that have to be chosen for every possible combination of atoms bound together.

The harmonic approximation is one of the severe limitations of this type of force fields. The two implications following immediately are that bonds can neither be formed nor broken and that the approximation to the exact PES is only good near the equilibrium distance r_0 where anharmonicity is not important.

Valence Angle Energy Terms

Analogous to the bonding energy term the valence angle contribution can be described by a harmonic potential. The function 2.8 is parameterized by the force constant κ_0 and the equilibrium angle θ_0 .

$$E_{\text{val}} = \sum_i \frac{1}{2} \kappa_0 (\theta_i - \theta_{i,0})^2 \quad (2.8)$$

The harmonic approximation for the angle contributions causes similar problems as were discussed for the bonding energy. The energy is ever increasing as the angle is increased to higher values. Looking at water as a simple example reveals the unphysical potential curves generated that way. It is well known that the water molecule has an equilibrium angle of 104.5° , the exact quantum chemical potential energy curve has another stationary point at 180° which is a maximum. The same potential curve calculated by a force field will yield a far too high energy for $\theta_{\text{HOH}} = 180^\circ$ and even worse, it has a cusp with an undefined gradient.

This means not only covalent reactions are excluded but angular rearrangements are somewhat limited too.

2. Theoretical Background

Dihedral Energy Terms

The general functional form of the dihedral energy is a linear combination of cosine terms with varying oscillation frequencies. The number of terms used can be changed depending on the molecular geometries that are to be simulated.

$$E_{\text{tors}} = \sum_i \sum_{n=1}^3 \frac{1}{2} V_{n,i} (1 + \cos(n\omega_i)) \quad (2.9)$$

The adjustable parameters V_n are the rotational barriers. In general three cosine terms for $n = 1, 2, 3$ in expression 2.9 are sufficient, as they can describe profiles for tetravalent species. Special cases like ferrocene for example would require higher terms with $n = 5$.

Coulomb Energy Terms

The *Coulomb* interactions are simply calculated by the *Coulomb* potential between two point charges.

$$E_{\text{coul}} = \sum_{i>j} \frac{e^2}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \quad (2.10)$$

The partial charges q_i and q_j in equation 2.10 are parameters assigned to the atom types of the i -th and j -th nucleus. Generally these parameters are found by fitting an electrostatic potential to quantum chemical electron densities for atoms in the respective chemical environment.

Van-der-Waals Energy Terms

The *Lennard-Jones* potential, sometimes also called 12-6 potential is the most common *Van-der-Waals* energy term. The potential in equation 2.11 is called this way because the competing repulsive and attractive terms are of 12th and 6th order respectively.

$$E_{\text{vdw}} = \sum_{i>j} 4\epsilon_{ij} \left(\left(\frac{r_{0,ij}}{r_{ij}} \right)^{12} - \left(\frac{r_{0,ij}}{r_{ij}} \right)^6 \right) \quad (2.11)$$

The parameter ϵ_{ij} is the atom pair specific depth of the *Lennard-Jones* potential well or *Van-der-Waals* dissociation energy, while the equilibrium distance is $\sqrt[6]{2}r_{0,ij}$.

Both, the *Coulomb* and the *Van-der-Waals* contributions, generally make up most of the the computational expense in molecular mechanics calculations. The quadratic scaling of the number of the unbound contributions compared to the linear increase in

2. Theoretical Background

bounded terms quickly outpaces them in terms of computing effort. For a molecule with 10 atoms roughly 33% of the energy terms are nonbonded, for 100 atoms this number already goes up to about 90% and even to 99% at 1000 atoms^[35]. In the limit of big systems, which is the intended application case of molecular mechanics, the scaling is therefore quadratic with the system size and almost exclusively dominated by nonbonded interactions. Of course, the scaling of the nonbonded interactions in practical applications is reduced by simplifications mentioned above, but still dominates the computational expense.

It is important to mention that in traditional force fields all the functions discussed above aren't parameterized per chemical element but per atom type.

An atom type is an element in a specific chemical environment, it depends on the element itself and the type of bonding it is involved in^[35]. This leads to a multitude of atom types per element for every possible functional group. For example the **OPLS-AA**^[36] parameter set as found in the potential library of the **TINKER**^[37] software package lists 906 atom types, 427 of which are carbon atom types. Therefore force field parameter sets rapidly grow into formidable databases. The atom types were invented to reflect for example the stark differences between equilibrium angles of sp³-, sp²- and sp-carbon of 109°, 120° and 180°. Thus the atom types can increase the accuracy of the force field, while at the same time simplify the parameterization for single atoms types.

To be fair it should also be noted that for most applications a small subset of atom types is sufficient to tackle a real life problem. Usually, less than 20 atom types with their respective parameters are used and the parameters are only weakly coupled. Therefore the parameterization of a force field for a specific process remains manageable.

However, the atom types pose one of the most severe limitations of those force fields. It is almost impossible to transition smoothly between atom types as chemical reactions occur. Hence the force fields are limited to cases where atom types remain unchanged, which excludes vast areas of chemistry. Reactive formalisms are aimed at eliminating that shortcoming while retaining the low computational demands of force fields.

2.2.2. Reactive Force Fields

In the previous section it was demonstrated that traditional force fields have severe limitations in reproducing covalent chemistry. While for example conformational changes in proteins or DNA^[38], solvation processes^[36] or stress on bulk materials can be simulated well, any reaction that includes the formation or cleavage of covalent bonds or a change in the chemical environment of atoms will push the traditional methods beyond their

2. Theoretical Background

limits. Reactive molecular mechanics methods are specifically designed to overcome these problems and at the same time retain most of the computational advantages gained over the quantum chemical methods.

If this sounds too good to come at no cost then that's because it doesn't. The additional flexibility of the reactive potentials is bought with a severe increase in their complexity. This introduces a manifold of difficulties for handling these formalisms, for example a much more challenging parameterization process or potential function instabilities to name just two.

To familiarize the reader with the topic, a short history of reactive formalisms and their similarities and differences will be given. This introduction is followed by a detailed analysis of the REAXFF potential functions and their parameters. The section is then concluded with a discussion of the challenges and opportunities posed by the REAXFF formalism.

Development of Reactive Force Fields

Today a wide variety of potentials exist which have successfully implemented the concept of reactivity. They may be classified into formalisms that are built on the concept of bond orders and those which are not, like for example the EVB^[39] method, QCT-FF^[40] or eFF^[41]. The bond order based method can further be divided into separated formalisms which have distinguishable n-body terms and integrated ones that include all n-body contributions via the bond order term^[42]. The *Finnis-Sinclair* model^[43], Embedded Atom Model EAM^[44] and Modified Embedded Atom Model MEAM^[45] come under the first class. The potentials by *Abell*^[46], *Tersoff*^[47], *Brenners* Reactive Empirical Bond Order (REBO) potential^[48] the REBO2^[49], the Adaptive Intermolecular Reactive Empirical Bond Order AIREBO method^[50] and REAXFF^[12] are examples for the second type.

One of the oldest approaches is the empirical valence bond or EVB method which was put forward by *Warshel* and *Weiss* in 1980^[39]. It uses multiple potential energy surfaces, for example of educt and product states, and a coupling matrix to ensure a smooth transition between them. The coupling matrix weights the potentials depending on geometrical parameters of the system or the energy difference between the PESs. Recently, the approach has been combined with *Grimme's* Quantum Mechanically Derived Force Field (QMDF) ^[51] to model reactions with a high accuracy^[52,53].

The separated formalism was developed around 1983 with the EAM potential and the *Finnis-Sinclair* model which is a special case of EAM^[42]. Both were published

2. Theoretical Background

independently from one another. Both formalisms lack angle-dependent energy terms and higher n-body terms. The introduction of an angle term in MEAM was meant to address the problem, but the potential still had problems with covalent systems.

The *Abell* model was published in 1985 and started a cascade of successor models based on it. In 1988 *Tersoff* extended *Abells* potential with an angle term. The additions made by *Brenner* in 1990 allowed for the formation of radicals. The resulting formalism can be found as *Brenner*-, *Abell-Tersoff-Brenner*- or REBO-potential in the literature^[48]. In 2002 REBO2 was published, it augmented the REBO formalism with torsional and *Lennard-Jones* contributions. Almost parallel to that development *Stuart* and coworkers published a potential called AIREBO in 2001 that was intended to model the transition between the REBO and the *Lennard-Jones* potential more accurately.

In 2001 *Adri van Duin* and coworkers also published their newly developed reactive formalism REAXFF^[12]. The core features were a completely parameterized bond order surface and dynamic charges by a self-consistent charge equilibration approach^[54,55]. The first version of REAXFF was only capable of modeling hydrocarbons^[12], but several extensions were made over the course of the next years to incorporate more effects in the potential^[56,57]. The current version was published in 2008 by *Chenoweth* and coworkers^[58]. *Liu* and coworkers added a term to account for *London* dispersion in 2011^[59], but this addition is not compatible with all REAXFF implementations and seldom used today.

The ReaxFF Formalism

As mentioned above, the first version of REAXFF was published in 2001^[12] as a hydrocarbon force field, but since then numerous modifications and additions were incorporated in the potential functions to mature it to its current version. Therefore all discussion will be based on the potential functions as published by *Chenoweth, van Duin* and *Goddard* in 2008^[58]. This is also the version that was implemented by *Aktulga*^[60] in the LAMMPS suite^[31].

In the current formulation the total energy of a system is comprised of 14 different contributions as seen in equation 2.12.

$$\begin{aligned} E_{\text{tot}} = & E_{\text{bond}} + E_{\text{lp}} + E_{\text{over}} + E_{\text{under}} + E_{\text{val}} + E_{\text{pen}} + E_{\text{coa}} + E_{\text{C2}} \\ & + E_{\text{triple}} + E_{\text{tors}} + E_{\text{conj}} + E_{\text{H-bond}} + E_{\text{vdw}} + E_{\text{coul}} \end{aligned} \quad (2.12)$$

2. Theoretical Background

The five terms E_{bond} , E_{val} , E_{tors} , E_{vdw} and E_{coul} are the same contributions that were discussed for traditional class I force fields before, but their functional forms differ fundamentally from the ones above. The reactive formalism furthermore accounts for overcoordination (E_{over}) and undercoordination (E_{under}), lone pair contributions (E_{lp}) and hydrogen bonding ($E_{\text{H-bond}}$). The remaining five terms are correction energies for rather specific situations. The penalty energy E_{pen} stabilizes allenes. The C_2 energy E_{C_2} corrects the erroneous behaviour of C_2 -molecules predicted by REAXFF. Another term for a specific molecule is the triple bond correction energy in carbon monoxide E_{triple} . The remaining two conjugation contributions E_{coa} and E_{conj} are used to capture the energetic effects of conjugated bonds. The most important terms of these will be discussed in detail in the following section.

Bond orders REAXFF is a bond order potential, therefore all bonded contributions mentioned above depend critically on the bond order between neighbouring atoms. The only two exceptions are of course the *Coulomb* and the *Van-der-Waals* interactions.

A fundamental assumption of *ReaxFF* is that the bond order of atom pairs can be obtained directly from the molecular geometry and depends solely on the internuclear distance between two atoms. This idea goes back to *Pauling* who coupled the bond order with the internuclear distance^[61]. At the very core of the ReaxFF potential is therefore the exponential relation between internuclear separation and the respective bond order in equation 2.13

$$\begin{aligned} BO'_{ij} &= BO'_{ij}{}^{\sigma} + BO'_{ij}{}^{\pi} + BO'_{ij}{}^{\pi\pi} \\ &= \exp\left(p_{\text{bo}1} \left(\frac{r_{ij}}{r_0^{\sigma}}\right)^{p_{\text{bo}2}}\right) + \exp\left(p_{\text{bo}3} \left(\frac{r_{ij}}{r_0^{\pi}}\right)^{p_{\text{bo}4}}\right) + \exp\left(p_{\text{bo}5} \left(\frac{r_{ij}}{r_0^{\pi\pi}}\right)^{p_{\text{bo}6}}\right) \end{aligned} \quad (2.13)$$

The uncorrected total bond order BO'_{ij} for atoms i and j is calculated as the sum of the single- $BO'_{ij}{}^{\sigma}$, the double- $BO'_{ij}{}^{\pi}$ and the triple bond order $BO'_{ij}{}^{\pi\pi}$. The primes indicate that the bond orders are uncorrected for now and will be adjusted later for over- and undercoordination effects. The bond orders are exponential functions of the internuclear distance r_{ij} and depend of three adjustable parameters each. The bond order of each bond type is restricted to $0 \leq BO'_{ij}{}^x \leq 1$ and strictly decreasing for all allowed values of the internuclear distance.

Starting from these initial values for the bond orders BO'_{ij} , corrections are applied. These correction rely on two overcoordination functions defined in equation 2.14.

2. Theoretical Background

$$\begin{aligned}\Delta'_i &= -Val_i + \sum_j^{N(i)} BO'_{ij} \\ \Delta_i^{\text{boc}} &= -Val_i^{\text{boc}} + \sum_j^{N(i)} BO'_{ij}\end{aligned}\tag{2.14}$$

Overcoordination Δ_i is the difference between the actual accumulated bond orders formed by an atom and all its neighbours and its number of electrons eligible for bonding. The parameter Val_i is the number of electrons that would engage in bonding per atom, it is for example 4 for carbon and 2 for oxygen. Note that this electron number is just a parameter as there are no explicit electrons in force field formalisms¹. The second overcoordination definition in equation 2.14 Δ_i^{boc} concerns atoms bearing lone pairs. In situations where lone pairs engage in bonding, the second overcoordination is employed to soften the energy contributions from other overcoordination terms. Take oxygen as an example: Oxygen usually bears two lone pairs which means the parameter Val_i^{boc} is 4. The hydronium ion is destabilized in the REAXFF potential because the valency of oxygen, which was 2, is exceeded. With the additional overcoordination definition it is possible to reduce this effect since now the lone pairs can correctly be used to form the bond.

Since the concept of bond orders is so fundamental to the REAXFF potential functions it is not advised to fiddle with the parameters Val_i and Val_i^{boc} when parameterizing the force field. Nonetheless there are published parameterizations using valency parameters which do not resemble the rationale described above. Although those parameterization can produce accurate results^[62], these parameters might not behave as expected, especially when parting from the systems in the reference data.

With the over coordinations at hand, an array of adjustments is applied to obtain the corrected bond orders.

$$\begin{aligned}BO_{ij}^\sigma &= BO'_{ij}^\sigma \cdot f_1(\Delta'_i, \Delta'_j) \cdot f_4(\Delta'_i, BO'_{ij}) \cdot f_5(\Delta'_j, BO'_{ij}) \\ BO_{ij}^\pi &= BO'_{ij}^\pi \cdot f_1(\Delta_i, \Delta'_j) \cdot f_1(\Delta'_i, \Delta'_j) \cdot f_4(\Delta_i^{\text{boc}}, BO'_{ij}) \cdot f_5(\Delta_j^{\text{boc}}, BO'_{ij}) \\ BO_{ij}^{\pi\pi} &= BO'_{ij}^{\pi\pi} \cdot f_1(\Delta'_i, \Delta'_j) \cdot f_1(\Delta'_i, \Delta'_j) \cdot f_4(\Delta_i^{\text{boc}}, BO'_{ij}) \cdot f_5(\Delta_j^{\text{boc}}, BO'_{ij}) \\ BO_{ij} &= BO_{ij}^\sigma + BO_{ij}^\pi + BO_{ij}^{\pi\pi}\end{aligned}\tag{2.15}$$

¹With the exception of *Goddard's* electron Force Field eFF^[41].

2. Theoretical Background

$$\begin{aligned}
 f_1(\Delta'_i, \Delta'_j) &= \frac{1}{2} \left(\frac{Val_i + f_2(\Delta'_i, \Delta'_j)}{Val_i + f_2(\Delta'_i, \Delta'_j) + f_3(\Delta'_i, \Delta'_j)} + \frac{Val_j + f_2(\Delta'_i, \Delta'_j)}{Val_j + f_2(\Delta'_i, \Delta'_j) + f_3(\Delta'_i, \Delta'_j)} \right) \\
 f_2(\Delta'_i, \Delta'_j) &= \exp(-p_{\text{boc1}} \cdot \Delta'_i) + \exp(-p_{\text{boc1}} \cdot \Delta'_j) \\
 f_3(\Delta'_i, \Delta'_j) &= -\frac{1}{p_{\text{boc2}}} \cdot \ln \left(\frac{1}{2} (\exp(-p_{\text{boc2}} \cdot \Delta'_i) + \exp(-p_{\text{boc2}} \cdot \Delta'_j)) \right)
 \end{aligned} \tag{2.16}$$

$$\begin{aligned}
 f_4(\Delta_i^{\text{boc}}, BO'_{ij}) &= \frac{1}{1 + \exp(-p_{\text{boc3}} \cdot (p_{\text{boc4}} \cdot (BO'_{ij})^2 - \Delta_i^{\text{boc}}) + p_{\text{boc5}})} \\
 f_5(\Delta_j^{\text{boc}}, BO'_{ij}) &= \frac{1}{1 + \exp(-p_{\text{boc3}} \cdot (p_{\text{boc4}} \cdot (BO'_{ij})^2 - \Delta_j^{\text{boc}}) + p_{\text{boc5}})}
 \end{aligned} \tag{2.17}$$

The first set of equations 2.15 scales the uncorrected bond orders with their correction factors. The resulting actual bond orders are summed to form the total bond order of an atom pair. The set of equations 2.16 contains the overcoordination corrections. To summarize the effect briefly it may be stated that the bond orders decay with increasing overcoordination Δ'_i . The remaining two equations (eq. 2.17), which are actually identical except for the argument Δ^{boc} , take care of bond order corrections by accounting for lone pairs.

The bond order functions are parameterized by a large number of empirical parameters. Two of the parameters used here, p_{boc1} and p_{boc2} , are general parameters unique to a parameter set. Another six parameters are atomspecific: r_0^σ , r_0^π , $r_0^{\pi\pi}$, p_{boc3} , p_{boc4} and p_{boc5} , thus they are set once per element. The remaining bond order parameters p_{bo} one through six are assigned for every possible bond type. Furthermore the radii r_0^x may be chosen for heteronuclear atom pairs as so-called off-diagonal terms to increase the accuracy of the parameterization.

An example parameter set containing the three elements carbon, hydrogen and oxygen therefore uses a minimum of 39 and up to 44 empirical parameters to build the bond orders. On top of that high dimensional hypersurface of geometry dependent bond order, it is now possible to construct and parameterize the REAXFF potential functions.

Bonding energy terms Three terms contribute to the bonding energy (eq. 2.18) for each bond type, i.e. bonding pair of atoms. The latter two, which govern bond order two and three, are the π and $\pi\pi$ bond orders scaled with a dissociation energy parameter.

2. Theoretical Background

The total energy contribution of both of them is negative and becomes zero as the bond order decays at large internuclear separations. The contribution of the σ bond has an additional exponential term that decays with the σ bond order BO_{ij}^σ .

$$E_{\text{bond}} = -D_e^\sigma \cdot BO_{ij}^\sigma \cdot \exp(p_{\text{be1}}(1 - (BO_{ij}^\sigma)^{p_{\text{be2}}})) - D_e^\pi \cdot BO_{ij}^\pi - D_e^{\pi\pi} \cdot BO_{ij}^{\pi\pi} \quad (2.18)$$

In stark contrast to the bonding potential of traditional force fields that were introduced in section 2.2.1, the REAXFF bonding energy is strictly increasing for all values $0 < r_{ij} < \infty$. The partitioning into an attractive and a repulsive part is common for bond order potentials since *Abell*^[42]. The bonding energy written in equation 2.18 is the attractive part of the bonding potential, while the repulsive part is mainly formed by the *Van-der-Waals* interactions that will be discussed later.

The count of adjustable parameters for the three-element potential example increases to 62 in the worst case scenario as twelve bond energy parameters and six bond order parameters are added. The parameters included at this point are used only to parameterize bonding energies. The comparison with the two parameters per bond, i.e. twelve for all possible bond types in the CHO example, necessary in traditional force field gives a rough idea of the complexity of reactive potentials.

Valence angle energy terms At the core of the valence angle energy term is the *Gaussian* function given in equation 2.19.

$$E_{\text{val}} = p_{\text{val1}} \cdot f_7(BO_{ij}) \cdot f_7(BO_{jk}) \cdot f_8(\Delta_j^{\text{boc}} (1 - \exp(-p_{\text{val2}}(\Theta_0(BO) - \Theta_{ijk})^2))) \quad (2.19)$$

This potential function has a negative contribution to the total energy at the equilibrium angle $\theta_{ijk} = \theta_0$ and becomes zero for large deviations of θ_{ijk} from θ_0 . The parameter p_{val1} is the depth of the potential well while p_{val2} controls its width.

The function $f_7(BO_{ij})$ (eq. 2.20) ensures that the energy contribution vanishes smoothly as the bond order between either atoms i and j or j and k decays to zero. Therefore no angle contributions are calculated for unbound atom pairs.

$$f_7(BO_{ij}) = 1 - \exp(-p_{\text{val3}} \cdot (BO_{ij})^{p_{\text{val4}}}) \quad (2.20)$$

Equation 2.20 accounts for overcoordination effects on the valence angle contribution.

2. Theoretical Background

$$f_8(\Delta_i^{\text{boc}}) = p_{\text{val}5} - (p_{\text{val}5} - 1) \cdot \frac{2 + \exp(p_{\text{val}6} \cdot \Delta_i^{\text{boc}})}{1 + \exp(p_{\text{val}6} \cdot \Delta_i^{\text{boc}}) + \exp(-p_{\text{val}7} \cdot \Delta_i^{\text{boc}})} \quad (2.21)$$

The empirical valence angle parameters three through seven have to be adjusted. The parameter $p_{\text{val}6}$ is a general parameter chosen once per parameter set, $p_{\text{val}3}$ and $p_{\text{val}5}$ are atom specific and the remaining parameters $p_{\text{val}4}$ and $p_{\text{val}7}$ are adjusted per atomic triple that forms an angle.

The last set of equations 2.22, 2.23 and 2.24 calculates the equilibrium angle θ_0 which depends on the molecular geometry, i.e. the coordination of the atoms involved.

$$SBO = \sum_{n=1}^{N(i)} (BO_{jn}^{\pi} + BO_{jn}^{\pi\pi}) + \left(1 - \prod_{n=1}^{N(i)} \exp(-BO_{jn}^8) \right) \cdot (-\Delta_j^{\text{boc}} - p_{\text{val}8} \cdot n_{\text{lp},j}) \quad (2.22)$$

$$SBO2 = \begin{cases} 0 & \text{if } SBO \leq 0 \\ SBO^{p_{\text{val}9}} & \text{if } 0 < SBO < 1 \\ 2 - (2 - SBO)^{p_{\text{val}9}} & \text{if } 1 < SBO < 2 \\ 2 & \text{if } SBO \geq 2 \end{cases} \quad (2.23)$$

$$\theta_0(BO) = \pi - \theta_{0,0} \cdot (1 - \exp(-p_{\text{val}10} \cdot (2 - SBO2))) \quad (2.24)$$

The valence angle parameters $p_{\text{val}8}$, $p_{\text{val}9}$ $p_{\text{val}10}$ are general parameters without a clear physical interpretation. $\theta_{0,0}$ is the equilibrium angle parameter and has to be adjusted for every angle.

The equations above are too convoluted for a detailed discussion to be fruitful, it is sufficient to recognize that the exponential term in equation 2.24 goes to unity with increasing π and $\pi\pi$ bond orders and is near zero for pure single bonds. As there are no atom types in REAXFF this function is needed to account for all possible equilibrium angles that geometries containing a certain element may exhibit. Carbon that was discussed above has three typical geometries, the tetrahedral angle in alkanes, the trigonal planar geometry in alkenes and a linear configuration in alkynes. With a parameter $\theta_{0,0}$ of roughly 70° all three angles can be reproduced^[58].

To compare the *ReaxFF* valence angle term to that of traditional force fields, the H₂O example from above shall be recalled. While the harmonic approximation shows

2. Theoretical Background

erroneous behaviour with a steep cusp for $\theta_{\text{HOH}} = 180^\circ$ the *Gaussian* function is able to describe the potential energy and gradients behaviour more correctly.

The valence angle term adds a large number of adjustable parameters to the CHO example considered before. With 96 additional parameters the number is increased to a total of 158.

Torsional angle terms The torsional angle term that models the rotational profiles in the REAXFF formalism looks similar to the profile discussed in section 2.2.1 for traditional force fields, but has the REAXFF-specific augmentations that were already seen before.

$$\begin{aligned}
 E_{\text{tors}} &= f_{10}(BO_{ij}, BO_{jk}, BO_{kl}) \cdot \sin(\theta_{ijk}) \cdot \sin(\theta_{jkl}) \cdot E'_{\text{tors}} \\
 E'_{\text{tors}} &= \frac{1}{2} (V_1(1 + \cos(\omega_{ijkl}) + F_{BO} \cdot V_2(1 - \cos(2\omega_{ijkl}) + V_3(1 + \cos(3\omega_{ijkl})) \quad (2.25) \\
 F_{BO} &= \exp\left(p_{\text{tor}1} (BO_{jk}^\pi - 1 + f_{11}(\Delta_j^{\text{boc}}, \Delta_k^{\text{boc}}))^2\right)
 \end{aligned}$$

The torsional barriers V_1, V_2 and V_3 have the same interpretation as in the traditional force fields. In contrast to the class I rotational profile, the torsional energy in equation 2.25 is scaled with a function that depends on the bond orders of the atom pairs involved and sine functions of the angle between three neighbouring atoms each. Once more the function $f_{10}(BO_{ij}, BO_{jk}, BO_{kl})$ in equation 2.26 is used to prevent torsional contributions to be calculated for unbound quadruples. The added sine functions will lead to vanishing torsional energy in linear geometries^[12].

$$\begin{aligned}
 f_{10}(BO_{ij}, BO_{jk}, BO_{kl}) &= (1 - \exp(-p_{\text{tor}2} \cdot BO_{ij})) \cdot (1 - \exp(-p_{\text{tor}2} \cdot BO_{jk})) \\
 &\quad \cdot (1 - \exp(-p_{\text{tor}2} \cdot BO_{kl})) \quad (2.26)
 \end{aligned}$$

Another addition is the exponential term F_{BO} in equation 2.25 which depends on the π bond order and the overcoordination of atoms j and k via the function $f_{11}(\Delta_j^{\text{boc}}, \Delta_k^{\text{boc}})$ (eq. 2.27). This function vanishes quickly when the central bond order BO_{jk} starts deviating from 2. It is therefore mainly effecting the cis-trans barrier in double bonded systems.

2. Theoretical Background

$$f_{11}(\Delta_j^{\text{boc}}, \Delta_k^{\text{boc}}) = \frac{2 + \exp(-p_{\text{tor}3} \cdot (\Delta_j^{\text{boc}} + \Delta_k^{\text{boc}}))}{1 + \exp(-p_{\text{tor}3} \cdot (\Delta_j^{\text{boc}} + \Delta_k^{\text{boc}})) + \exp(p_{\text{tor}4} \cdot (\Delta_j^{\text{boc}} + \Delta_k^{\text{boc}}))} \quad (2.27)$$

The torsional potential introduces another seven adjustable parameters to the force field. Four of them are assigned per atom quadruple, the other three are general parameters. Unfortunately the number of possible combinations for these quadruples grows with the fourth power of the number of parameterized elements. Thus the number of empirical parameters becomes unmanageable at four elements if all of them are to be considered. To keep the number of empirical parameters for the torsional potential in check, *van Duin* utilizes wildcards in the potential. It is possible to specify the parameters for a certain quadruple of four atoms, or just to specify the two central atoms and use dummy atoms for the first and the fourth one. This means when all possible torsional parameters are defined for CHO there are another 207 free parameters, in a best case scenario where wildcard are utilized torsionals only 27 of them remain. The total number of parameters in the best case scenario is therefore 185. Usually this number is a little higher as some torsionals are desired to have higher accuracy than others and therefore use specific parameterizations instead of the wildcard option. In the disulfide parameterization discussed in section 4.3, five explicit torsionals were used besides the wildcards.

Coulomb energy terms The *Coulomb* interaction in the REAXFF formalism is calculated according to equation 2.28.

$$E_{\text{Coul}} = \text{Tap}(r_{ij}) \cdot C \cdot \frac{q_i q_j}{\left[r_{ij}^3 + \left(\frac{1}{\gamma_{ij}} \right)^3 \right]^{\frac{1}{3}}} \quad (2.28)$$

The function $\text{Tap}(r_{ij})$ (eq. 2.29) is a seventh order polynomial taper function which ensures that the nonbonded interaction goes smoothly to zero. This way no singularities occur in the first derivative of the potential due to the cutoff distance for nonbonding interactions at r_{cut} .

$$\text{Tap}(r_{ij}) = \frac{20}{r_{\text{cut}}^7} r_{ij}^7 - \frac{70}{r_{\text{cut}}^6} r_{ij}^6 + \frac{84}{r_{\text{cut}}^5} r_{ij}^5 - \frac{35}{r_{\text{cut}}^4} r_{ij}^4 + 1 \quad (2.29)$$

The denominator of equation 2.28 is for the shielding of the *Coulomb* potential at low distances. The shielding effectively prohibits excessively high repulsion energies,

2. Theoretical Background

sometimes called *Coulomb* catastrophe, between point charges near each other. The shielding parameter γ is element specific. The pairwise parameter γ_{ij} is derived from the atomic parameters as their geometric mean.

Due to the flexibility of the reactive force field, which should allow smooth transitions of atoms between different chemical environments, the point charges q_i of the atoms can not be constant parameters. In contrast to the traditional force fields the charges are obtained depending on the molecular geometry with a self-consistent method instead. These dynamic charges are needed because partial charges may change drastically during chemical reactions when atoms are transitioning through various chemical environments. The dynamic update of charges during MD simulations is of course superior to the description by parametric partial charges, but the self-consistent scheme is one of the most expensive step in terms of computational resources.

The self-consistent QEq procedure was a performance enhancement by *Rappe* and *Goddard*^[55] of *Mortier*'s^[54] EEM method. The basic assumption for this method is that the partial charges within any molecule adjust themselves such that the electronegativity of all individual atoms becomes equal. This condition yields $n - 1$ expressions for the electronegativities of an n atom system.

$$\chi_i = (\chi_{i,0} + \Delta\chi_i) + 2(\eta_{i,0} + \Delta\eta_i)q_i + \sum_{j \neq i} \frac{q_j}{r_{ij}} \quad (2.30)$$

$$\chi_i = \chi_j = \chi_k \dots \chi_n$$

In equation 2.30 χ is the electronegativity while η is the electronic hardness. The electronegativity is the atoms tendency to attract electron density while the electronic hardness is a measure for the tendency to retain electron density. The initial values χ_0 and η_0 are those for the free atoms. The changes $\Delta\chi$ and $\Delta\eta$ occur due to the chemical environments. Due to the constraint that all χ_i have to be equal in equation 2.30 there are only $n - 1$ equations to calculate the n partial charges. The last equation needed to form a complete set of linear equations is the condition that the sum of all partial charges q_i has to equal the total molecular charge Q_M (eq. 2.31).

$$Q_M = \sum_i^n q_i \quad (2.31)$$

The adjustable parameters for the *Coulomb* potential and the QEq method, respectively, are the atomic parameters for electronegativity χ_0 and the electronic hardness η_0

2. Theoretical Background

as well as the shielding parameter γ . This adds another nine parameters to the three-atom example which currently totals to 194 parameters. It should be mentioned that parameters χ and η are published for almost the entire periodic table of elements and do not require much optimization, if any.

Van-der-Waals energy terms Similar to the *Coulomb* energy, the *Van-der-Waals* energy (eq. 2.32) is tapered to avoid discontinuities at the cutoff distance. It is also shielded at small internuclear separations by the shielding factor γ_{vdw} in the function $f_{13}(r_{ij})$. This prevents excessive repulsions energies.

$$E_{\text{vdw}} = \text{Tap}(r_{ij}) \cdot D_{ij} \cdot \left[\exp \left(\alpha_{ij} \left(1 - \frac{f_{13}(r_{ij})}{r_{\text{vdw}}} \right) \right) - 2 \exp \left(\frac{1}{2} \alpha_{ij} \left(1 - \frac{f_{13}(r_{ij})}{r_{\text{vdw}}} \right) \right) \right]$$

$$f_{13}(r_{ij}) = \left[r_{ij}^{p_{\text{vdw}1}} + \left(\frac{1}{\gamma_{\text{vdw}}} \right)^{p_{\text{vdw}1}} \right]^{\frac{1}{p_{\text{vdw}1}}}$$
(2.32)

In contrast to the 12-6 potential mentioned for the class I force fields, REAXFF uses a *Morse* potential for the long range interactions.

However, let a point made above be stressed once more. The repulsive part of the *Van-der-Waals* potential forms the repulsive part of the two-body term, i.e. the total bonding energy. Therefore an accurate parameterization of the *Van-der-Waals* contributions is critical even for systems that only have neglectable long-range interactions.

The potential depends on four adjustable parameters: D_{ij} , α_{ij} , r_{vdw} and γ_{vdw} . The shielding radius γ_{vdw} is a atom specific parameter, while the other three may be assigned per atom or pairwise. The latter is recommended to ensure the flexibility needed for different bond types. With that another 21 parameters are needed for the CHO example case. The total number is therefore 215.

Remaining energy terms The remaining nine energy terms that are listed in equation 2.12 are only of small relevance for this thesis and are therefore not discussed here in detail. Detailed information on their functional form and use for the REAXFF potential can be found in *van Duins* original publication^[12], as well as the supporting information of the 2008 paper^[58].

A Perspective on ReaxFF

The previous section gave a very detailed overview of the REAXFF potential functions. The discussion shall be concluded by a comparison between reactive force fields and their non-reactive counterparts.

In terms of flexibility the reactive potential has the clear lead. While traditional force fields only allow for conformational rearrangements, REAXFF can simulate the whole diversity of covalent chemistry. The restrictions of atom types and predefined bonding motifs do not apply to reactive potentials.

Of course the more versatile potential functions are more complex and demand higher computation time to be evaluated. Furthermore the self-consistent QEq method to derive the atomic partial charges takes its toll. Therefore non-reactive force fields are computationally much cheaper, making them the favorable choice if no covalent processes are involved. Reactive formalisms however are themselves considerably cheaper than DFT or semiempirical quantum chemistry methods. They are routinely used for system sizes varying from a few thousand to over a million atoms and simulations on the nanosecond scale.

It may have occurred already to the reader that the previous section kept count of the adjustable parameters that have to be optimized to yield a reactive force field for three atoms. Here the real drawback of reactive potentials becomes evident. Even for small systems a staggering amount of empirical parameters is needed, some of them with shady or no physical interpretations at all. Compared to the parameters in non-reactive potentials they may be coupled in very intricate fashion. To overcome this difficulty, performant global optimization techniques and high-quality reference data for every relevant region of the PES is required. The following sections will therefore cover global optimization techniques, specifically global optimization using evolutionary algorithms (section 2.3) and high level ab-initio methods capable of generating the needed reference data (section 2.4).

2.3. Global Optimization of Parameters Sets

If chemical reactions are to be investigated by molecular mechanics simulations, the empirical parameters of the force field employed need to be adjusted appropriately. The usual course of action is to consider a limited subset of information about a system and find a parameter set which reproduces the data as accurately as possible.

This set of information is often referred to as training or reference set. It can be com-

2. Theoretical Background

prised of any data that may be obtained about a chemical system either experimentally or by theoretical methods. The reference data for force field parameterizations typically include molecular equilibrium geometries, potential energies, transition state structures, partial charges and gradient information. Depending on the aim of the parameterization, more involved properties of a system may be fitted, for example volumetric mass densities, rate constants, heats of formation, bulk moduli and unit cell parameters. Any data may be used for a reference as long as rules can be defined to obtain them from molecular mechanics and quantum chemical methods.

The best empirical parameters for a set of potential functions is then found via an optimization procedure. The set of parameters that yields the lowest deviation from the reference set is considered optimal.

In case of the REAXFF optimization in the presented in this thesis, the deviation from the reference set is calculated as a sum of weighted and squared differences between reference values and empirical potential results. Equation 2.33 thus is a measure for the quality of a parameter set^[12].

$$F_{\text{obj}}(x_i) = \sum_i \frac{(x_i - x_{i,\text{ref}})^2}{w_i^2} \quad (2.33)$$

In optimization problems $F_{\text{obj}}(x_i)$ in equation 2.33 is called an objective function^[63]. The optimization task is to find the set of parameters that yields the minimal value for equation 2.33, where zero means perfect representation of the reference set. The weighting factors w_i scale the contributions of reference items to the error sum. They are used to account for different units of the data or its importance for the parameterization.

In the previous section it was already mentioned that for a reactive force field the number of empirical parameters which need adjustment can easily be in the hundreds. Not only is the search space high-dimensional, it also has an extremely challenging structure. Figure 2.6 showcases a two-dimensional cut through the objective function of an SiOH parameter set. The function has a lot of characteristics associated with hard optimization problems: Ruggedness, deceptiveness, discontinuities and multimodality. Another quality, which is not obvious in the figure, is the coupling of the parameters^[1].

The already difficult optimization problem has one further complication to it. Any gradient-based optimization method will inevitably fail when operating on $F_{\text{obj}}(x_i)$. The objective function contains properties like molecular geometries or partial charges which need to converge to a self-consistent state in order to return an objective value. Numerous regions of the search space correspond to parameter settings that will yield

2. Theoretical Background

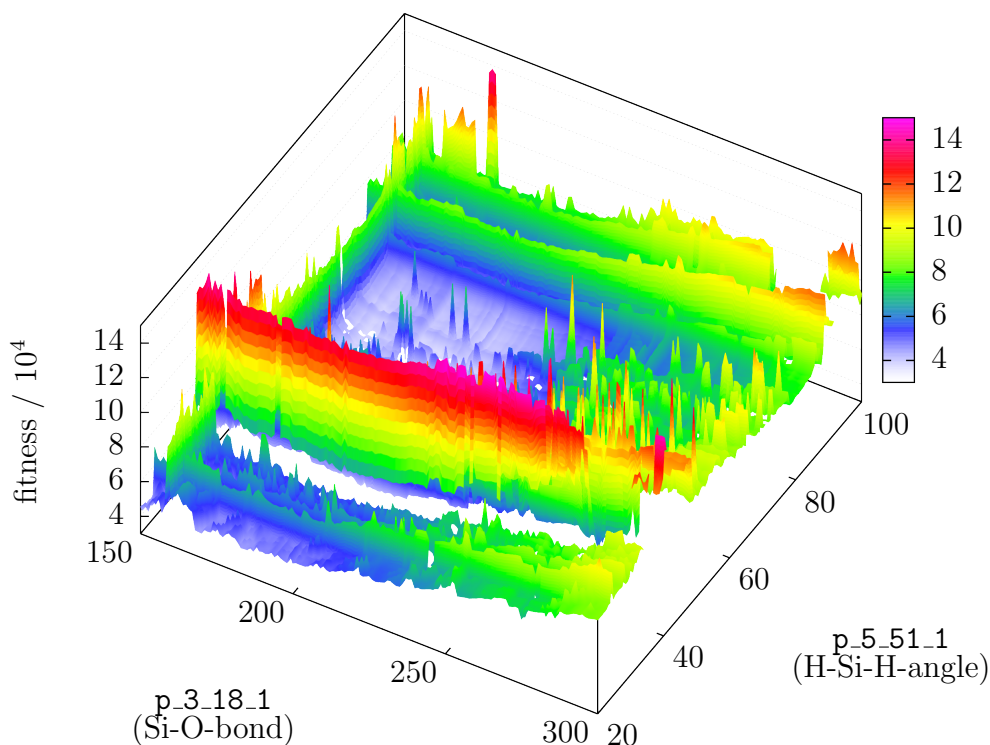


Figure 2.6.: Two-dimensional cut through the objective function of a 67 dimensional SiOH optimization problem. The objective function value is plotted over a bond energy parameter and an equilibrium angle parameter. Figure reprinted with permission. Copyright 2015 Wiley^[1].

unconverged geometries or charges for molecules in the reference set. In those regions, objective value and gradient are undefined. Additionally, as shown in figure 2.6, there are discontinuities where the gradient is undefined. Another restriction to the calculation of the gradient is a practical one. The dependence of the objective value on the parameters is too convoluted to derive an analytical gradient and a numerical one would be computationally very expensive.

The global optimization problem outlined above presumably falls in the class of NP-hard problems. This means there is no known deterministic algorithm to solve it in polynomial time. Since the exponential time scaling of any deterministic optimizer effectively prohibits solving the problem for all but the smallest, and in practice mostly uninteresting cases, non-deterministic strategies are needed. Heuristics are such strategies that offer a probability to find good solutions in polynomial time. In contrast to deterministic algorithms these good solutions are not guaranteed to be the globally optimal solution. In practice this is usually of no concern as there is no need for the exact globally optimal solution as long as high-quality approximate solutions exist on

the objective function.

Initially, optimization of the REAXFF parameters were done by the quasi-local SOPPE routine^[12]. While many workgroups continue to use local optimization methods, global optimization schemes are applied by some groups to deal with increasingly complex and high-dimensional parameterization problems. Metaheuristics have been successfully used to optimize parameters, for example Metropolis Monte Carlo (MMC) methods with a simulated annealing (SA) approach^[64] or multi objective evolutionary strategies (MOES)^[65]. This thesis is focussed on the global optimization of force fields with Single Objective Evolutionary Algorithms (SOEAs or just EAs) which will be introduced in the following section. Following the discussion of the global optimization procedure, the quantum chemical methods to obtain the needed reference data will be covered.

2.3.1. Evolutionary Algorithms

To quote the excellent textbook on global optimization by Thomas Weise^[63]:

“Evolutionary algorithms (EAs) are population-based metaheuristic optimization algorithms that use biology-inspired mechanisms like mutation, crossover, natural selection, and survival of the fittest in order to refine a set of solution candidates iteratively.”

At this point it is crucial to note that none of the algorithmic performance of EAs arise just because they are bio-inspired. For all the progress made by nature-inspired technology, it is not a constructive approach to value terminology and close mimicking of nature over rigorous mathematical models. In fact EAs are often hybridized or specialized to perform better for a certain class of optimization problems. This reduces their similarities to biological evolution, which are already debatable, even further. To keep terminology transferable between various bio-inspired concepts and non-inspired optimizations, and discourage the idea of algorithms being good just by being bio-inspired, it is advised to keep it formal and restrict the use of inspired terminology to the necessary minimum.

Weise's definition of the term evolutionary algorithm lacks explanatory power as to what all the terms mean and how EAs really work. For a better introduction, the fundamental concept of a basic EA is described first and then limitations and possible improvements are discussed.

The General Evolutionary Algorithm

The evolutionary algorithm is started off with the initialization of a population P . The population P is a collection of solution candidates I_a that are generated by a nullary initialization operator. The solution candidate is a vector to a point in the search space, its entries are the adjustable empirical parameters. During the initialization usually all vector entries are set to random values restricted by the search space boundaries. For the optimization the EA operates on the vector entries and varies them in search of superior solution candidates. In case of the REAXFF reparameterization these vector entries are the empirical parameters.

The evaluation step assigns every solution candidate an objective value. This is simply done by forwarding the vector to the objective function (eq. 2.33) and by saving the return value $F_{\text{obj}}(I_a)$.

During the fitness assignment, a fitness value f_a for each solution candidate is computed. In the most simple case the fitness value is just the result of a user-defined fitness function acting on the objective value $F_{\text{fit}}(F_{\text{obj}}(I_a)) = f_a$. There are of course far more sophisticated methods to calculate fitness values some of which will be discussed later.

The reproduction step is initiated by choosing good solution candidates with a probability according to their fitness for the mating process. For a simple example the probability may just be calculated by a normalization of each solution candidate's fitness f_a with respect to the sum of all fitnesses in the population P . A random number between 0 and 1 is then generated and a solution candidate chosen accordingly. This random draw results in the best candidates to be most likely chosen for mating, and is termed roulette wheel selection. This is again just one of many different methods to choose solution candidates.

The two solution candidates selected this way are subjected to unary and binary operations, usually less formally referred to as mutation and crossover. The binary operator, depicted on the left side in figure 2.7, generates two new solution candidates by cutting two vectors at a random point and exchanging the cut parts between them. The resulting vector may then be mutated by an unary operator that varies candidate vector entries, i.e. parameter values, randomly.

With a new population formed by the solution candidates obtained from the mating, the cycle begins again at the evaluation step.

Eventually some kind of user-defined termination criterion is reached, then the optimization stops and returns the solution candidates. The absence of a reliable measure for the progress of the EA makes the termination problematic. The objective value

2. Theoretical Background

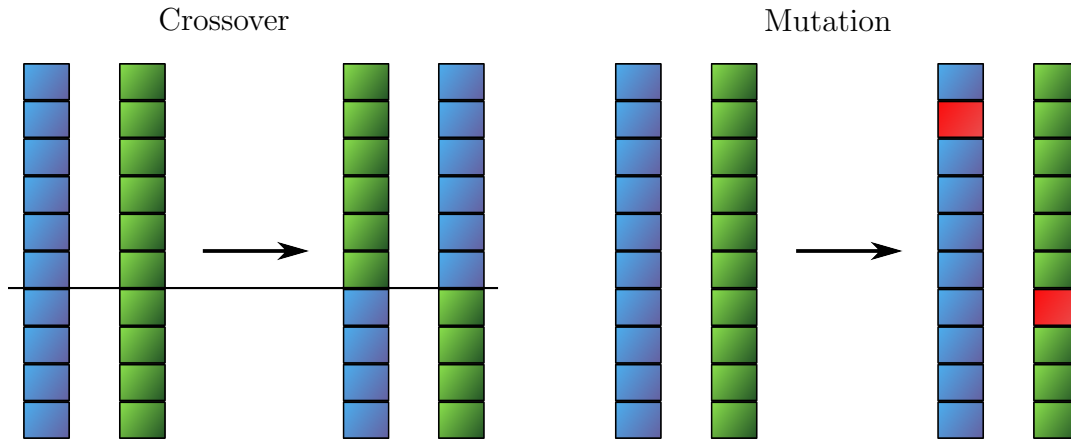


Figure 2.7.: Schemes for the binary crossover operator (left) and the unary mutation operator (right) that are used during the reproduction step. The binary operator cuts two solution candidates at a random point and the resulting partial vectors are exchanged to form new solution candidates. In an unary mutation step, a random, usually small, amount of parameters is altered to new random values.

may still improve erratically after many iterations without a change of the best solution candidate. Typically EAs terminate after a certain amount of iterations or computing time.

The big selling point of the EA described here is the applicability of the algorithm to any optimization problem regardless of the solution and the objective function. It is therefore possible to apply EAs to any problem with a suitable objective function in a black-box fashion^[63]. However, the algorithm described here may perform rather poorly in an actual optimization. As stated by the "No-Free-Lunch-Theorems", there is no truly universal fast optimization algorithm that can perform well on all problems^[66]. Simple improvements, some of which are discussed below, can enhance the performance of EAs greatly while making them less general applicable at the same time.

Shortcomings of the Simple Evolutionary Algorithm and Improvements

It was already mentioned that the design described above has some flaws which diminish the efficiency of the algorithm in certain situations. The problems will be discussed in the following paragraphs and solution strategies will be provided. Since all optimizations done in this thesis almost exclusively used the OGOLEM^[67,68] suite, the operators implemented there will be briefly introduced.

As for many other global optimization algorithms, premature convergence is a problem

2. Theoretical Background

for EAs when operating on deceptive multimodal objective functions. Some good initial solutions may lead to a swift proliferation of the respective partial solutions through the population and a quick descent into an area with promising local minima which are outside the funnel of the global optimal solution. There the algorithm may get trapped for a long time. All following iterations are likely to exploit these local minima further while other regions of the search space remain unexplored. Although the search operations of the EA used in this thesis are complete in theory this has limited practical relevance. Completeness in this sense means that it is possible to reach any point within the search space from any other point with the search operations provided to the algorithm^[63]. In practice the probability to find a promising solution candidate that is also accepted into the population vanishes as soon as the population starts converging towards better objective values.

The problem of premature convergence basically means that the population is losing diversity faster than it converges towards the global optimum^[63]. Countermeasures that may be employed are multiple restarts of the optimization or enforcing diversity by including diversity measures in the fitness assignment or in the selection process.

Starting an EA multiple times, each time with a randomly initiated population, may indeed yield the optimal result, but depending on the structure of the objective function, the chance to end up with the optimal result will diminish with increasing dimensionality of the problem. The probability to find a solution candidate in the vicinity of the global optimum increases linearly with the number of restarts or the population size. Since a multimodal objective function has its number of local minima scaling exponentially with its dimensionality, this number will quickly outpace the amount of local minima that can be compensated for by generating a higher number of random starting points, i.e. restarts. Another complication with multiple restarts is a property of the REAXFF objective function associated with the keyword domino-convergence^[63,69]. Some empirical parameters have a large influence on the objective value, while other parameters only give small contributions. Since the effect of the small contributions can not be estimated in the beginning of the optimization, where the objective value improves in large steps, multiple restarts yield no improvement for the less contributing parameters.

The more promising way is keeping the population diverse by active measures during the optimization. A population P obviously becomes less diverse when a large number of similar solution candidates are present. A very straightforward way to prevent similar solution vectors from accumulating in P is to penalize the fitness of solution candidates with many others near them. Since the empirical parameters of the candidate form a

2. Theoretical Background

vector pointing to a solution in the search space, the distance of solution candidates may for example be defined by the Euclidean distance between them. An appropriate sharing function is then used to reduce the candidates fitness according to number and distance of other nearby solutions. This concept was introduced by *Goldberg* and *Richardson* in 1987^[70].

In the current implementation of OGOLEM the diversity can be enforced by the niching concept. A niche is a confined area of the search space which is then only allowed to be inhabited by a user-defined small portion of the total population. This niching is not implemented via rescaled fitness values as used by *Goldberg* and *Richardson* but as a second acceptance criterion besides the objective value.

Another difficulty comes with the operators themselves. In an EA that is not hybridized with a local optimization for intermediate solutions, the binary single-point crossover operator described above can only generate solution candidates on the vertices of a hypercuboid in the searchspace. These vertices are defined by the empirical parameter values of both solution candidates used for the crossover. Crossover operators can have problems with ordering, parameter coupling and excessive dimensionality which may lead to situations where a single random cutting point yields undesired results or slows down the convergence process dramatically. There are even scenarios where the operation on the parameter vector has to be abandoned to successfully solve a problem^[71].

Boon and bane of evolutionary algorithms is that there is no unique correct way to do unary or binary operations on the solution candidates. A binary operator is not confined to have a single random cutting point, multiple points may be used and their position constrained in any way that suits the user's demands. Furthermore ternary or higher order operators can be used to generate new solution vector and thus enhance the optimization performance.

To address the problem of solutions being just vertices of an n dimensional hypercuboid where n is the number of optimizable parameters, it is possible to depart from the crossover concept. Offspring parameters can be mixed as weighted averages from parent vectors. Another possibility is to resort to propagation algorithms that generate new solution candidates as linear combinations of the ones chosen for mating.

All these techniques may improve the performance of the algorithm, but it is sometimes unclear which operators will perform well for the problem at hand.

For the REAXFF parameterization with OGOLEM three binary operators and two unary operators are at the user's disposal. Two of the binary operators are single- and n -point crossovers while the third operator allows weighted averaging of parent

2. Theoretical Background

parameters. One of the unary operators reinitializes random parameters to a value in the search space, while the other generates Gaussian-distributed random values for the parameters with an adjustable width and center position for the Gaussian^[1,72]. It is furthermore possible to combine all operators freely to generate customized operators. The probability to invoke unary or binary operator can be chosen, as well as coefficients for every operation described above. These coefficients determine the probability for the operator to be invoked.

It was already mentioned that it is unclear which operator performs best in a given situation. To a certain extent this may be estimated by reasoning. An unary operator that reinitializes a solution candidate has great exploratory power whilst operators that vary solutions only a little are used for exploitation. The operators offered by OGOLEM are a flexible toolkit for all these situations. However the reasoning has its limitations and explicit testing in search of the optimal operator settings has to be done. For the REAXFF parameterization such testing was done and discussed in our publication^[1].

With the keyword of exploitation a further challenge arises. While the operators of an EA are very useful for finding wells and dips of an objective function, they are ill-suited to exploit the local optimum within a well area.

To compensate for this shortcoming EAs are often hybridized with local optimization routines. In case of geometry optimizations the local relaxation may be done in every iteration, what greatly increases the convergence speed. If a local optimization can not be afforded or is not desired in every iteration, like it is the case for the *ReaxFF* parameterization, local optimization can be done at the end of the calculation or in certain long intervals throughout the run.

The OGOLEM implementation allows for various different local optimization schemes in the REAXFF parameterization. It was found that the most efficient and convenient way for local optimization of empirical parameters was to manually curate promising solution candidates at the end of a global optimization run and relax them locally. This local optimization was done with a greedy hillclimb algorithm which was emulated by using the Gaussian unary operator with a narrow distribution around the current parameter values.

A final challenge that has to be considered, particularly in the recent years, is the parallel scalability of the EA. The generational concept possesses a serial bottleneck by its nature and has therefore limited parallel scaling. This hampers the treatment of large-scale optimization problems on highly parallel computing clusters like the HLRN^[73,74].

The bottleneck exists in operations like the selection for mating or fitness sorting to

2. Theoretical Background

check whether or not a solution candidate is fit to be accepted in the population. These operations can not be well parallelized what leads to idle times of workers when they are waiting for other evaluations to finish or the sorting to complete.

Hartke and *Bandow* addressed the problem of parallel scaling by abandoning the EA’s generational paradigm in favor of a pool-based steady-state EA^[73]. The whole population is now stored in a pool which is handled by a master process. The master hands out selected pairs, or packs of pairs, to worker processes that will handle unary and binary operations as well as evaluation of the candidates. The solution candidates are then returned to the master and inserted into to pool if all criteria, i.e. objective value and possible diversity checks, are met. The worst solutions are discarded in the process to keep the pool at a constant size.

The pool concept is also implemented in the OGOLEM code. It offers excellent parallel scaling for the optimization of molecular geometries as well as REAXFF parameters^[1,68,74].

2.4. Wave Function Methods

In computational chemistry solutions to *Schrödinger’s* equation (eq. 2.34)^[75] need to be approximated numerically.

$$\hat{H}\Psi = E\Psi \quad (2.34)$$

The solutions Ψ are the eigenfunctions of the electronic Hamiltonian \hat{H} within the *Born-Oppenheimer* picture. The following section will be used to briefly introduce the basic approximation, i.e. the *Hartree-Fock* method, and discuss qualitative errors of the resulting wave function and how to fix them. The improved methods beyond *Hartree-Fock* are needed to compute the reference data which includes covalent chemistry and are therefore required for accurate REAXFF parameterizations. Most of the following information is a review of the standard textbooks on the matter^[35,76–79], if references besides those have been used they are cited explicitly.

2.4.1. Basic Approximation and the Hartree-Fock Method

In the *Born-Oppenheimer* approximation the total molecular hamiltonian \hat{H} is simplified to an electronic hamiltonian that contains the nuclear coordinates just as parameters. In atomic units the electronic hamiltonian takes the form shown in equation 2.35.

2. Theoretical Background

$$\hat{H}_e = -\frac{1}{2} \sum_i^{n_{el}} \nabla_i^2 + \sum_i^{n_{el}} \sum_{j>i}^{n_{el}} \frac{1}{r_{ij}} - \sum_i^{n_{el}} \sum_a^{n_{nu}} \frac{Z_a}{r_{ia}} + \sum_a^{n_{nu}} \sum_{b>a}^{n_{nu}} \frac{Z_a Z_b}{r_{ab}} \quad (2.35)$$

The first term governs the kinetic energy of the electrons. It is followed by terms for the repulsive potential between electron pairs and the attractive potential between nuclei and electrons. The last operator calculates the repulsive contributions of pairs of nuclei. It is constant for a given molecular geometry and therefore just an offset parameter added to the total energy in the *Born-Oppenheimer* picture.

As no analytical solution may be obtained for the equations above for more than three particles, numerical procedures are required to approximate the molecular hamiltonian and its respective eigenfunctions. In the self-consistent scheme devised by *Douglas Rayner Hartree* and *Wladimir Alexandrowitsch Fock* (hence the name *Hartree-Fock-Method*) the wave function takes the form of a *Slater-determinant* (eq.2.36)

$$\Psi_{SD} = \begin{vmatrix} \phi_1(1) & \phi_2(1) & \dots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \dots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \dots & \phi_N(N) \end{vmatrix} \quad (2.36)$$

In the *Slater-determinant* the molecular spin-orbitals ϕ_i are arranged along the columns and depending on different electronic spin coordinates arranged along the rows. This form of the wave function ensures antisymmetry upon label exchange between two fermions.

In the numerical approximation the Hamiltonian takes the form of the *Fock-operator* (eq.2.37). It is assumed that the electron pair repulsion can be approximated as the average effect all remaining electrons have on one. This description lacks contributions stemming from the exact more instantaneous interaction between electrons or correlation, which is of profound importance for high-quality wave functions.

$$\hat{F}_i = \hat{h}_i + \sum_{j \neq i}^{n_{el}} (\hat{J}_j - \hat{K}_j) \quad (2.37)$$

In the final set of *Hartree-Fock-equations* (eq.2.38) using the molecular spin-orbitals and the *Fock-operator*, the orbitals are stationary with respect to their variation.

$$\hat{F}_i \phi_i = \epsilon_i \phi_i \quad (2.38)$$

2. Theoretical Background

These are the canonical *Hartree-Fock*-orbitals and their respective energies. The MOs found in this procedure yield the best approximation to the wave function a single *Slater* determinant can give. However, some small but nonetheless very important contributions to the energy are still missing. This missing energy, the difference between the exact result and the *Hartree-Fock*-energy, is generally called electron correlation energy.

$$E_{corr} = E_{exact} - E_{HF} \quad (2.39)$$

It is possible to distinguish two forms of correlation contributions. The dynamic correlation consists of the contributions from the direct and instantaneous electron-electron interaction that is neglected in the *Hartree-Fock*-Method due to its characteristic mean field approximation. Static correlation becomes important in situations where *Born-Oppenheimer* surfaces of chemically inequivalent electronic configurations become near degenerate and interact with each other. When this is the case several electronic configurations, or *Slater* determinants, become important for the electronic wave function. Examples for such degeneracies occur in conical intersections or along bond dissociation curves.

2.4.2. Dynamic Correlation and Møller-Plesset Perturbation Theory

A generally used method to include small corrections to an already known solution is perturbation theory. In quantum mechanics *Rayleigh-Schrödinger* perturbation theory (RSPT) is applied to correct for small errors in the eigenvalues and eigenfunctions. First the hamiltonian is partitioned in a zeroth-order hamiltonian and a perturbation.

$$\hat{H} = \hat{H}_0 + \lambda \hat{V} \quad (2.40)$$

The perturbation \hat{V} is added to the unperturbed hamiltonian \hat{H}_0 and can be switched on and off by the parameter λ . To apply the perturbation to the wavefunction and eigenvalues they are expanded in a power series with respect to λ .

$$\Psi_n = \Psi_n^{(0)} + \lambda \Psi_n^{(1)} + \dots + \lambda^m \Psi_n^{(m)} = \sum_{i=0}^m \lambda^i \Psi_n^{(i)} \quad (2.41)$$

$$E_n = E_n^{(0)} + \lambda E_n^{(1)} + \dots + \lambda^m E_n^{(m)} = \sum_{i=0}^m \lambda^i E_n^{(i)} \quad (2.42)$$

When the expressions 2.40, 2.41 and 2.42 are inserted into the *Schrödinger* equation

2. Theoretical Background

2.34 it transforms to expression 2.43.

$$\left(\hat{H}_0 + \lambda\hat{V}\right) \sum_{i=0}^m \lambda^i \Psi_n^{(i)} = \left(\sum_{i=0}^m \lambda^i \Psi_n^{(i)}\right) \left(\sum_{i=0}^m \lambda^i E_n^{(i)}\right) \quad (2.43)$$

Expanding the equation above and collecting equal powers of λ yields conditional equations for perturbative corrections of arbitrary order. The equations up to second order are given in eq. 2.44.

$$\begin{aligned} \hat{H}_0 \Psi_n^{(0)} &= E_n^{(0)} \Psi_n^{(0)} \\ \hat{H}_0 \Psi_n^{(1)} + \hat{V} \Psi_n^{(0)} &= E_n^{(0)} \Psi_n^{(1)} + E_n^{(1)} \Psi_n^{(0)} \\ \hat{H}_0 \Psi_n^{(2)} + \hat{V} \Psi_n^{(1)} &= E_n^{(0)} \Psi_n^{(2)} + E_n^{(1)} \Psi_n^{(1)} + E_n^{(2)} \Psi_n^{(0)} \end{aligned} \quad (2.44)$$

Multiplying from the left with $\Psi_n^{(0)*}$ and using the relation $\langle \Psi_n^{(0)} | \Psi_n^{(m)} \rangle = \delta_{m0}$ yields expressions for the perturbation energies.

$$\begin{aligned} E_n^{(0)} &= \langle \Psi_n^{(0)} | \hat{H}_0 | \Psi_n^{(0)} \rangle \\ E_n^{(1)} &= \langle \Psi_n^{(0)} | \hat{V} | \Psi_n^{(0)} \rangle \\ E_n^{(2)} &= \langle \Psi_n^{(0)} | \hat{V} | \Psi_n^{(1)} \rangle \end{aligned} \quad (2.45)$$

As the eigenstates of the hamiltonian form a complete set the first order wave function needed for the second-order energy contribution may be expanded in these eigenfunctions.

$$\Psi_n^{(1)} = \sum_i c_i \Psi_i^{(0)} ; c_i = \frac{\langle \Psi_i^{(0)} | \hat{V} | \Psi_n^{(0)} \rangle}{E_n^{(0)} - E_i^{(0)}} \quad (2.46)$$

Møller-Plesset perturbation theory (MP), named after *Christian Møller* and *Milton S. Plesset* who published the concept in 1934, is derived by applying the RSPT to the HF method. Although other formalisms may arise too, depending on the choice zeroth-order hamiltonian and the perturbation, this text focusses on MP theory. More specifically only second-order *Møller-Plesset* perturbation theory (MP2) is discussed, as higher-order perturbations quickly become computationally unfeasible while providing only small further corrections to the wave function or the energy.

The unperturbed hamiltonian H_0 used in the MP2 theory is a sum of *Fock* operators for every electron in the system. The perturbative correction to the operator is to include

2. Theoretical Background

the direct electron-electron interaction and to get rid of the error introduced by averaging over the effect of the electrons.

$$\begin{aligned}\hat{H}_0 &= \sum_i \hat{F}_i = \sum_i \hat{h}_i + \hat{V}_i^{HF} \\ \hat{V} &= \hat{V}_{ee} - 2\langle V_{ee} \rangle = \sum_{i<j} \frac{1}{r_{ij}} - \hat{V}_i^{HF}\end{aligned}\tag{2.47}$$

Using the optimized *Slater* determinant $\Psi_n^{(0)}$ as zeroth-order function and the operators above, it can be shown that the sum of zeroth and first-order energies is the HF ground state energy. To further include correction that entail the dynamic correlation energy of the system, the second-order energy is needed. Equation 2.45 shows that the second-order energy is dependent of the first-order corrected eigenfunction. In the MP2 formalism the first-order eigenfunction is expanded in a basis of doubly excited *Slater* determinants (eq. 2.48).

$$\begin{aligned}\Psi_n^{(1)} &= \sum_{a>b;r>s} c_{abrs} \langle \Psi_{ab,n}^{rs} \rangle \\ c_{abrs} &= \sum_{a>b;r>s} \frac{\langle \Psi_{ab,n}^{rs} | \Psi_n^{(0)} \rangle}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s}\end{aligned}\tag{2.48}$$

The final MP2 energy may then be calculated according to equation 2.49.

$$E_n^{(2)} = \sum_{a>b;r>s} \frac{|\langle \Psi_{ab,n}^{rs} | \hat{V} | \Psi_n^{(0)} \rangle|^2}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s}\tag{2.49}$$

From this energy expression it follows immediately that MP2 results become very unreliable if excited determinants exist that degenerate with the *Hartree-Fock* ground state, as the denominator approaches zero and energy contributions become large. Unfortunately this is almost inevitably the case in dissociations.

2.4.3. Density Functional Theory

The most commonly used post *Hartree-Fock* method by far that takes dynamic correlation into account is density functional theory (DFT). The approach in DFT is to include the correlation energy via an exchange-correlation functional that depends on the elec-

2. Theoretical Background

tron density in a system. The *Hartree-Fock* exchange is either completely replaced in LDA, GGA and meta-GGA functionals or partially retained in hybrid and double-hybrid functionals. For details refer to the literature on the topic^[35,76].

However, as modern DFT methods are all built on HF wave functions, the fundamental flaw of the missing static correlation remains. To date no widely used approach to include static correlation in the DFT formalism exists. DFT is therefore not eligible to build reliable wave functions in multireference regions of the PES. Therefore no DFT data is used for the reference set in CMC parameterization.

2.4.4. Static Correlation and Multireference Methods

Besides dynamic correlation the second correlation contribution is static correlation which is sometimes also called near-degeneracy effect. It occurs in situations where different electronic states approach each other in energy. This is the case in avoided crossings along dissociation curves or with near-degenerate biradicalic states in the dissociation limit.

Multireference methods are approaches to include static correlation effects into the wave function. As in the perturbation theory, correlation is introduced by means of excited *Slater* determinants. The basic idea of the configuration interaction (CI) approach is to expand the exact wave function in a basis of excited *Slater* determinants. It should be noted that the series expansion shown here uses unaltered *Hartree-Fock*-orbitals and is therefore still a single reference wavefunction.

$$\Psi_n = c_{HF}\Psi_{HF} + \sum_S c_S\Psi_S + \sum_D c_D\Psi_D \cdots \sum_N c_N\Psi_N \quad (2.50)$$

The unaltered *Hartree-Fock* solution is Ψ_{HF} in the subsequent singly, doubly or n-tuply excited determinants one, two or N electrons are elevated to virtual orbitals, respectively. In a CI calculation the CI coefficients c_i are optimized variationally while the coefficients determining the HF spin orbitals are left unaltered. The orbital coefficients are not visible in equation 2.50 but hidden in the *Slater* determinants instead.

Note that the full CI expansion in 2.50 when used with a complete basis set yields the exact solution for the wavefunction, including the full correlation either dynamic or static. But the full expansion is computationally unfeasible for all but the smallest systems. Truncated methods like CISD, which only accounts for singly and doubly excited determinants, may be less resource demanding but may lack the configurations essential for static correlation contributions.

2. Theoretical Background

This single reference approach may be improved to a multireference ansatz by using more than one primary *Slater* determinant to generate excited configurations.

$$\Psi_n = \sum_i^R \left(c_{i,0} \Psi_{i,0} + \sum_S c_{i,S} \Psi_{i,S} + \sum_D c_{i,D} \Psi_{i,D} \right) \quad (2.51)$$

In the expansion equation 2.51 uses singly and doubly excited determinants built from R reference configurations $\Psi_{i,0}$. This ansatz is sometimes referred to as MRCISD. While now all important configurations can be considered the wave function has at least two fundamental flaws. First, the shape of *Hartree-Fock*-orbitals may change drastically when electrons are excited into virtual orbitals. Second, it is not trivial to choose reference configurations as there is no reliable measure for the importance of determinants a priori.

The first of the problems is addressed by the multiconfigurational self consistent field (MCSCF) approach. The MCSCF procedure doesn't only optimize the CI coefficients of the reference determinants but the MO coefficients as well.

The complete active space self consistent field method (CASSCF) is a widely used successor of MCSCF that also tackles the second problem with a more chemical approach to reference choice. CASSCF utilizes a definite set of orbitals chosen prior to the calculation. The orbitals chosen are called active space (AS) and should be comprised of all orbitals that are relevant for the chemical process investigated. In case of a single bond dissociation the active space may consist of a σ orbital and the corresponding σ^* orbital. Such an AS would be denoted (2,2) and the complete calculation is called CASSCF(2,2). The first number gives the number of electrons in the active space, the second one counts the spatial orbitals. Within such an active space a full CI expansion is done and the orbital coefficients of each configuration are optimized with respect to one electronic state.

The state averaged CASSCF method or SA-CASSCF allows to treat multiple root states, i.e. excited states of defined spin multiplicity, with an equal degree of accuracy. To have molecular orbitals optimized to represent different electronic states equally well is of interest in situations where multiple near degenerate electronic states contribute considerably to the wave function character. This is the case for example at avoided crossings, conical intersections or intersection seams.

Despite the power of the CASSCF method to retrieve large parts of the static correlation, the approach is ill-suited to account for the relevant dynamic correlation. Of course the CASSCF is able to also yield the full dynamical correlation, as full CI is able,

but the active space has to be excessively large to do so. This means another layer of theory on top of CASSCF is needed for an accurate method.

2.4.5. The CASPT2 Multireference Perturbation Theory

Above the basics of MP2 and MCSCF or more specifically CASSCF were discussed. It was stated that perturbative treatment of the wave function can retrieve major parts of the dynamic correlation energy. CASSCF can account for the static correlation present in near-degeneracy regions of the PES. To have a truly universal tool both approaches were combined by *Roos* and coworkers to build a complete active space second-order perturbation theory (CASPT2) method.

Without any details to the intricacies of the CASPT2 method it should be mentioned that the CASPT2 wave function is a second-order perturbation treatment on top of an optimized CASSCF reference function. The combined correlation energy of both approaches accounts for big parts of the total correlation energy of a system. The use of suitable basis functions like correlation consistent triple zeta basis sets and a proper active space leads to correct relative energetics between arbitrary points on the PES investigated. For further detailed information on the MCSCF and CASPT2 methods the reader is redirected to the relevant literature on this subject^[35,79].

3. Triazole Mechanochemistry

3.1. Scope of the Project

Within the SFB677 the collaborative subproject A05 was concerned with mechanochemistry at a single-molecular level. Systems of interest were molecules that can undergo force-induced conformational changes or have predetermined molecular breaking points (PBPs) which may be exploited in AFM experiments. These so called mechanophores were approached from multiple angles, including synthesis, setting up and conducting AFM experiments and theoretical investigation.^[11] The author was solely concerned with theoretical simulations throughout the project.

In their now retracted paper “Unclicking the Click: Mechanically Facilitated 1,3-Dipolar Cycloreversions”, *Brantley, Wiggins* and *Bielauski* did claim to have reversed the “click”-reaction^[80], also known as *Huisgen’s* reaction or azide-alkyne cycloaddition (AAC), for 1,2,3-triazoles by exposing them to shearing forces exerted by ultrasonic treatment^[8]. The then standing motivation for the project was to use macrocyclic systems in a single-molecule force spectroscopy (SMFS) setup to see whether the cycloreversion can be induced and detected in single molecules using an AFM tip.

The macrocyclic system introduced in the publication^[2] is a novel experimental approach to pinpoint single-molecular reactions. A so-called safety line is utilized to bridge the PBP, i.e. the 1,2,3-triazole moiety, of the mechanophore. This bridging allows to unambiguously identify a fission of the PBP, since when a break occurs in the bridged region a subsequent rupture of the safety line will be detected as a second peak on the force extension curve. Furthermore the experimentalists are able to discriminate between situations where two fissions occur in a single molecule and cases where multiple ruptures can be attributed to several mechanophores suspended in the AFM. This is because the slope of the force extension curves depends on number of suspended mechanophores and their attachment angle.

Throughout the project two challenges occurred. First, there were few double rupture events detected and the resulting scarcity of data made quantitative interpretation of

3. Triazole Mechanochemistry

the results difficult. Second, the inability to differentiate between a cycloreversion and a single bond fission in the PBP moiety lead to the question whether positive experimental results can really be attributed to a cycloreversion into alkyne and azide products. So despite the possibility to pinpoint the rupture to a very confined region of the molecule there was no certainty to what process really was observed.

Throughout the early project phase and the publication process the main concern for the theoretical investigations was to settle whether the experimental findings are backed by theory. Although theoretical and experimental results compared well, it became more and more apparent that the simple explanation of a cycloreversion reaction is not supported by the theoretical results. These suspicions became much more evident as a letter of concern and ultimately the retraction hit the original publication by *Bielawski* and coworkers^[8]. The theoretical investigation therefore underwent a shift from validating the experiment to actually find out which processes can explain the findings from *Schütze's* experiments.

3.2. Publication: Pinpointing Mechanochemical Bond Rupture by Embedding the Mechanophore into a Macrocycle.^[2]

Contribution to the paper:

- Setup of small model systems.
- Setup and execution of ab-initio simulations on different levels of theory.
- Interpretation of static data and comparison with experimental results.

Full text reprinted with permission. Copyright 2015 Wiley.

Pinpointing Mechanochemical Bond Rupture by Embedding the Mechanophore into a Macrocycle**

Doreen Schütze, Katharina Holz, Julian Müller, Martin K. Beyer,* Ulrich Lüning,* and Bernd Hartke*

Dedicated to Professor Hermann E. Gaub on the occasion of his 60th birthday

Abstract: Mechanophores contain a mechanically labile bond that can be broken by an external mechanical force. Quantitative measurement and control of the applied force is possible through atomic force microscopy (AFM). A macrocycle was synthesized that contains both the mechanophore and an aliphatic chain that acts as a “safety line” upon bond breaking. This ring-opening mechanophore unit is linked to poly(ethylene glycol) spacers, which allow investigation by single molecule force spectroscopy. The length increase upon rupture of the mechanophore was measured and compared with quantum chemical calculations.

A wide variety of mechanochemical reactions have been demonstrated through the deliberate incorporation of mechanophores^[1–3] into long polymers,^[3,4] with the aim of developing mechanoresponsive materials. Several cycloreversion mechanophores have been reported.^[5–7] A 1,2,3-triazole moiety embedded in poly(methyl acrylate) appeared to undergo mechanochemical cycloreversion,^[6,7] which would seem to demonstrate that “click” chemistry is mechanically reversible. However, the validity of these experimental data is currently under debate.^[8]

In order to address the putative mechanochemical cycloreversion of 1,2,3-triazoles with a different experimental

approach, we designed a ring-opening mechanophore that allows the investigation of the mechanochemical activation of a 1,2,3-triazole on the single-molecule level. In contrast to simple bond-breaking mechanophores, which lead to polymer rupture at a defined site, ring-opening mechanophores lead to an elongation of the polymer upon activation. The ring-opening mechanophore with the largest known elongation, at 0.4 nm, is a bicyclo[3.2.0]heptane.^[9] Herein, we report the synthesis of a ring-opening mechanophore with an elongation of more than 1.0 nm. The elongation was directly measured by single-molecule force spectroscopy (SMFS)^[10,11–14] and compared to quantum chemical calculations.

The design of the mechanophore was inspired by the work of Fernandez and co-workers, who used an engineered protein to study the force-dependence of bimolecular disulfide reduction by SMFS.^[15] In our ring-opening mechanophore **14**, the bond to be cleaved is in the shorter branch of the macrocycle, namely that containing the triazole moiety. The longer branch is an alkyl chain, which constitutes the “safety line”. The carboxylic acid end groups allow us to incorporate the mechanophore in between poly(ethylene glycol) (PEG) spacers, which are required for SMFS.

Triazole **14** was synthesized in 10 steps, starting with protection of the carboxylic acid of **1** as the corresponding methyl ester.^[16] Etherification of **2** with hex-5-en-1-ol (**3**) through a Mitsunobu reaction gave **4**. For the reduction of the nitro group, stannous dichloride dihydrate was used and the desired amine **5** was obtained. Amine **5** is needed to produce both the azide **7** and the iodide **8**.

First, aniline **5** was transformed into the corresponding diazonium salt **6** by adding hydrochloric acid and sodium nitrite at 0 °C. After the addition of sodium azide, product **7** was obtained. The diazonium salt **6** could also be used to produce iodide **8** through a Sandmeyer analogous reaction. To convert **8** into the alkyne **10**, a Sonogashira coupling with trimethylsilyl acetylene was performed. The silylated alkyne **9** was deprotected and **10** was obtained (Scheme 1).

Starting from alkyne **10** and azide **7**, triazole **11** was produced in a copper-catalyzed [3+2] cycloaddition (click reaction) by employing microwave irradiation. The safety line was introduced through ring-closing metathesis and macrocyclic alkene **12** was isolated. Hydrogenation of the double bond was carried out with platinum(IV) oxide and hydrogen to yield the saturated macrocycle **13**. Ester cleavage as the final step led to dicarboxylic acid **14** (Scheme 2 and the Supporting Information).

[*] Dipl.-Chem. D. Schütze,^[†] M. Sc. J. Müller,^[†] Prof. Dr. B. Hartke

Institut für Physikalische Chemie
Christian-Albrechts-Universität zu Kiel
Olshausenstraße 40, 24098 Kiel (Germany)
E-mail: hartke@pctc.uni-kiel.de
Homepage: <http://ravel.pctc.uni-kiel.de/>

M. Sc. K. Holz,^[†] Prof. Dr. U. Lüning
Otto-Diels-Institut für Organische Chemie
Christian-Albrechts-Universität zu Kiel
Olshausenstraße 40, 24098 Kiel (Germany)
E-mail: luening@oc.uni-kiel.de
Homepage: <http://www.luening.otto-diels-institut.de/de>

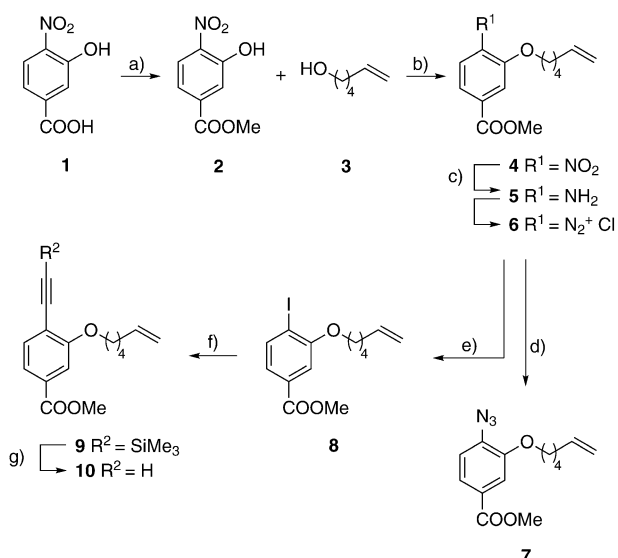
Prof. Dr. M. K. Beyer
Institut für Ionenphysik und Angewandte Physik
Leopold-Franzens-Universität Innsbruck
Technikerstraße 25, 6020 Innsbruck (Austria)
E-mail: martin.beyer@uibk.ac.at
Homepage: <http://www.uibk.ac.at/ionen-angewandte-physik/>

[†] These authors contributed equally to this work.

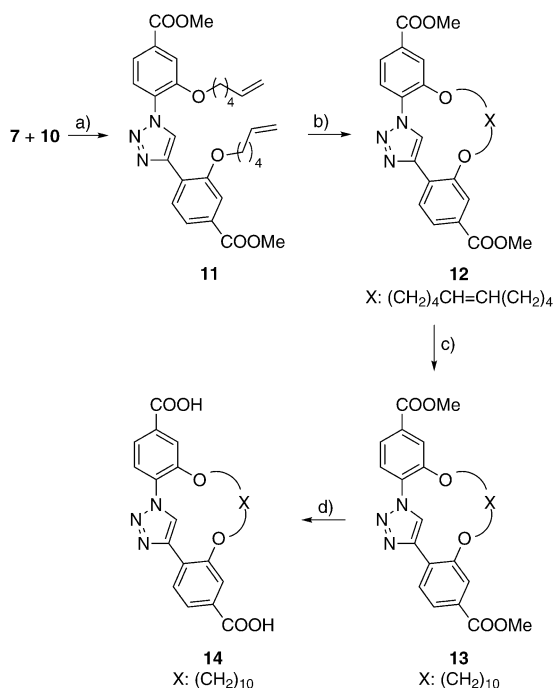
[**] Financial support from the Deutsche Forschungsgemeinschaft in the SFB 677: “Function by Switching” is gratefully acknowledged. Compound **2** was provided by Isabel Köhl.



Supporting information for this article is available on the WWW under <http://dx.doi.org/10.1002/anie.201409691>.

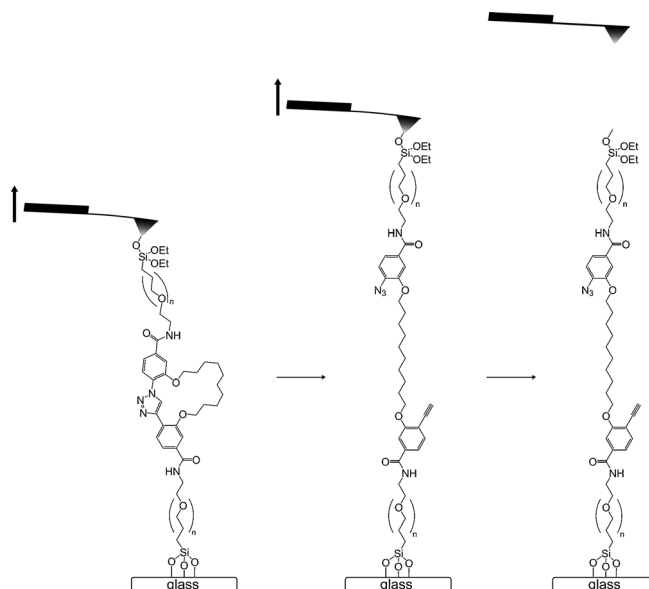


Scheme 1. Synthesis of azide **7** and alkyne **10**: a) MeOH, H₂SO₄, 24 h, reflux, 99%; b) 1. THF, PPh₃, 2. DIAD, 30 min, 0 °C, 3. 22 h, RT, 83%; c) EtOH, SnCl₂·2H₂O, AcOH, 1 h, 75 °C, 87%; d) 1. H₂O, HCl, 0 °C, 2. NaNO₂, 20 min, 0 °C, 3. NaN₃, 30 min, 0 °C, 57%; e) 1. Acetone, HCl, 0 °C, 2. NaNO₂, 2 h, 0 °C, 3. KI, 30 min, 0 °C, 15 min, 80 °C, 81%; f) Me₃SiC≡CH, THF, Pd(PPh₃)₂Cl₂, CuI, NEt₃, 20 h, RT, 74%; g) CHCl₃, Bu₄NF, 16 h, RT, 95%.



Scheme 2. Synthesis of macrocycle **14** from azide **7** and alkyne **10**: a) MeCN, CuI, EtNiPr₂, MW, 10 min, 100 °C, 120 W, 71%; b) CH₂Cl₂, Grubbs' catalyst 1st gen., 36 h, RT, 77%; c) CHCl₃, PtO₂, H₂, 24 h, RT, 97%; d) THF, MeOH, H₂O, LiOH·H₂O, 1. 5 min, 50 °C, 2. 15 h, RT, 96%.

To investigate the mechanochemical ring-opening behavior of **14**, the molecule was covalently attached via an amide bond to a PEG chain, which was covalently attached through silane anchors to a glass substrate (Scheme 3). SMFS was



Scheme 3. **14** is covalently anchored between two PEG chains, which in turn are covalently attached between a glass substrate and a Si₃N₄ cantilever. Upon stretching of the molecule, a double rupture event is observed if bond rupture occurs first in the triazole branch of **14** (cycloreversion shown), and a characteristic length increase is measured by AFM.

employed in fly-fishing mode.^[17] The PEG-silanzed cantilever repeatedly approached the glass substrate, with the tip continuously covered by the solution. In less than 10% of the approaches, a second amide bond was formed between the amine end group of the PEG and the second carboxylic acid of **14**. The force–extension curve, where extension refers to the piezo element of the cantilever, exhibited the characteristic shape of a stretched PEG molecule^[18] (Figure 1).

In the vast majority of successful attachments, only a single rupture event was observed, which is associated with rupture of the silane surface anchor.^[11–14,19] In about 5%

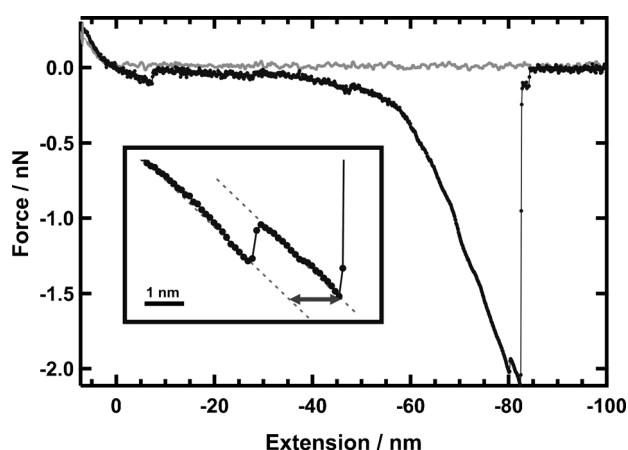


Figure 1. Force–extension curve with one of the rare double rupture events assigned to the rupture of **14** (curve 3 in Table 1). The inset shows that the slope of the force–extension curve is the same before and after the rupture. The length change is read from the figure as the horizontal displacement of the force–extension curve.

of the recorded force–extension curves, however, two rupture events were observed. These double rupture events were due either to the mechanochemical ring opening of macrocycle **14** or to the attachment of two polymer chains. In the case of mechanochemical ring opening, the same polymer chain was stretched before and after the first rupture event, thus resulting in identical slopes of the curve (Figure 1). If multiple polymer chains were attached, the slope of the force–extension curve changed after the first rupture, and these curves were discarded. After close examination of all of the force–extension curves with double ruptures, only three curves remained in which the slope was the same before and after the first rupture event. In these curves, the length increase is measured as described in Figure 1. The results are summarized in Table 1, with a conservative uncertainty of ± 0.2 nm for the length change, based on the uncertainty of positioning the parallel fit lines in the force–extension curves.

Table 1: Rupture force and elongation measured from force–extension curves featuring a double rupture event.

Force–extension curve	Rupture Force [nN]	Elongation Δx (exp) [nm]	Elongation Δx (COGEF) [nm]
1	1.11 ± 0.01	1.2 ± 0.2	1.01 ± 0.20
2	1.21 ± 0.02	1.2 ± 0.2	1.01 ± 0.20
3	2.05 ± 0.03	1.4 ± 0.2	1.05 ± 0.20

According to Bielawski and co-workers,^[6,7] this length increase should be assigned to a mechanochemical retro-click reaction of macrocycle **14**. However, bond ruptures between the triazole unit and its phenyl anchors, with the triazole ring remaining intact, would lead to the same AFM response. These events thus cannot be distinguished here, and the single-molecule nature of the experiment precludes the use of standard spectroscopic techniques to further differentiate between the possible products.

The observation of only three ring-opening events in several thousand force–extension curves clearly shows that the aryl–triazole–aryl region is mechanically stronger than the silane surface anchor.^[11] In force-ramp experiments, the rupture forces are scattered over a range of more than 1 nN,^[12] so it is entirely reasonable that in rare cases the mechanically stronger bond breaks first. With the safety line concept described here, we can unambiguously identify these events and measure the rupture force. Unfortunately, the events were so rare in this case that a quantitative statistical analysis was not possible.

Among the quantum-chemical methods available to describe covalent mechanochemistry,^[2,20,21] the constrained geometries simulate external force (COGEF) method^[20] is ideally suited to describe AFM experiments performed in force-ramp mode. In this study, COGEF calculations were used to determine the expected elongation associated with mechanochemical ring-opening of **14** and subsequent stretching of the $(\text{CH}_2)_{10}$ safety line. Since only the length difference before and after mechanochemical ring-opening was of interest here, only the force-induced structural deformability of the initial and final states was modelled.

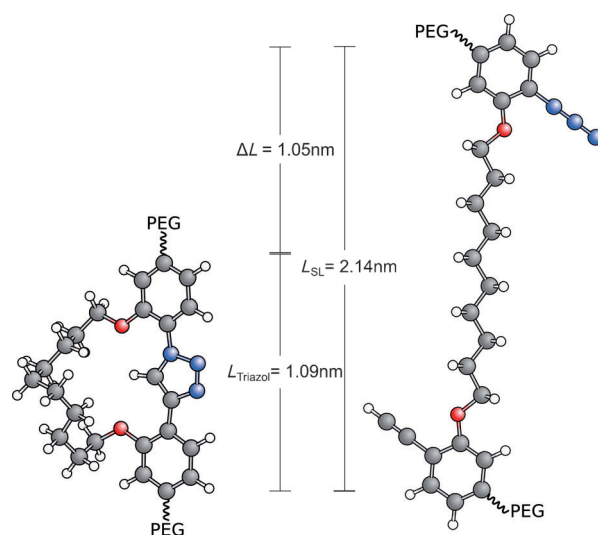


Figure 2. COGEF calculation of the length increase resulting from mechanochemical ring opening of macrocycle **14**.

The initial state was represented by two phenyl rings linked by a 1,2,3-triazole; the final state by two phenyl rings linked by the safety line (Figure 2). The PEG chains were not included in the calculations. For both configurations, a series of relaxed scans was performed. The pulling forces were obtained from the first derivatives of the energy–distance curves. From the resulting force–distance curves, the length difference between the initial and final configurations, taken at the experimental force value, yielded our theoretical estimate for the length change upon mechanochemical ring-opening. This theoretical estimate is 1.05 ± 0.20 nm for a force of 2.05 nN, which is slightly shorter than the experimental value. However, there are several effects that contribute to significant uncertainties. The conditions for the calculations were set as under vacuum at 0 K, while the experiments were performed in solution at room temperature. Rupture of the C–N bond instead of triazole cycloreversion will slightly change the geometry of the aryl groups. The high force of 2.05 nN applied to the bond deforms the binding potentials and increases their anharmonicity, which in turn leads to thermal expansion of the PEG chain. The effect of the lever arms was also shown to be significant.^[22] These effects justify a conservative estimate of ± 0.20 nm for the uncertainty.

The described combination of tailor-made mechanophore synthesis, AFM experiments, and quantum chemical calculations shows that arbitrary bonds can be embedded in a macrocycle and selectively addressed through an external mechanical force applied through PEG linkers. Even very few rupture events of the mechanophore can be unambiguously identified through the characteristic length increase together with the unchanged slope of the force–extension curve before and after the rupture event. With the present molecular design, we cannot determine whether it was really mechanochemically induced retro-click reactions of the 1,2,3-triazole ring that took place or merely bond ruptures next to it. We can state, however, that the force required to induce either of the two reactions is in the nN region. The present technique opens a wide range of possibilities for the design of

mechanophores. By changing the length of the safety line in **14**, ring-opening mechanophores with arbitrary elongation can be synthesized. By replacing the 1,2,3-triazole linkage, for example with cyclobutane or disulfide, the response of the mechanophore can be finely tuned to a specific range of mechanical strain. These two degrees of freedom make a wide variety of mechanophores accessible for the design of functional materials.

Received: October 2, 2014

Published online: January 22, 2015

Keywords: cycloreversion · DFT calculations · macrocycles · mechanophores · single-molecule force spectroscopy

- [1] a) M. K. Beyer, H. Clausen-Schaumann, *Chem. Rev.* **2005**, *105*, 2921–2948; b) C. R. Hickenboth, J. S. Moore, S. R. White, N. R. Sottos, J. Baudry, S. R. Wilson, *Nature* **2007**, *446*, 423–427; c) M. M. Caruso, D. A. Davis, Q. Shen, S. A. Odom, N. R. Sottos, S. R. White, J. S. Moore, *Chem. Rev.* **2009**, *109*, 5755–5798; d) J. N. Brantley, K. M. Wiggins, C. W. Bielawski, *Angew. Chem. Int. Ed.* **2013**, *52*, 3806–3808; *Angew. Chem.* **2013**, *125*, 3894–3896.
- [2] J. Ribas-Arino, D. Marx, *Chem. Rev.* **2012**, *112*, 5412–5487.
- [3] J. N. Brantley, K. M. Wiggins, C. W. Bielawski, *Polym. Int.* **2013**, *62*, 2–12.
- [4] Z. S. Kean, S. L. Craig, *Polymer* **2012**, *53*, 1035–1048.
- [5] a) M. J. Kryger, M. T. Ong, S. A. Odom, N. R. Sottos, S. R. White, T. J. Martinez, J. S. Moore, *J. Am. Chem. Soc.* **2010**, *132*, 4558–4559; b) M. J. Kryger, A. M. Munaretto, J. S. Moore, *J. Am. Chem. Soc.* **2011**, *133*, 18992–18998; c) H. M. Klukovich, Z. S. Kean, S. T. Iacono, S. L. Craig, *J. Am. Chem. Soc.* **2011**, *133*, 17882–17888; d) K. M. Wiggins, J. A. Syrett, D. M. Haddleton, C. W. Bielawski, *J. Am. Chem. Soc.* **2011**, *133*, 7180–7189; e) J. N. Brantley, S. S. M. Konda, D. E. Makarov, C. W. Bielawski, *J. Am. Chem. Soc.* **2012**, *134*, 9882–9885.
- [6] J. N. Brantley, K. M. Wiggins, C. W. Bielawski, *Science* **2011**, *333*, 1606–1609.
- [7] Y. Lin, Q. Wang, *Angew. Chem. Int. Ed.* **2012**, *51*, 2006–2007; *Angew. Chem.* **2012**, *124*, 2046–2047.
- [8] M. McNutt, *Science* **2014**, *344*, 1460.
- [9] Z. S. Kean, A. L. Black Ramirez, Y. Yan, S. L. Craig, *J. Am. Chem. Soc.* **2012**, *134*, 12939–12942.
- [10] a) E. L. Florin, V. T. Moy, H. E. Gaub, *Science* **1994**, *264*, 415–417; b) P. Schwaderer, E. Funk, F. Achenbach, J. Weis, C. Bräuchle, J. Michaelis, *Langmuir* **2008**, *24*, 1343–1349; c) J. Liang, J. M. Fernández, *J. Am. Chem. Soc.* **2011**, *133*, 3528–3534.
- [11] M. Grandbois, M. Beyer, M. Rief, H. Clausen-Schaumann, H. E. Gaub, *Science* **1999**, *283*, 1727–1730.
- [12] S. W. Schmidt, M. K. Beyer, H. Clausen-Schaumann, *J. Am. Chem. Soc.* **2008**, *130*, 3664–3668.
- [13] S. W. Schmidt, A. Kersch, M. K. Beyer, H. Clausen-Schaumann, *Phys. Chem. Chem. Phys.* **2011**, *13*, 5994–5999.
- [14] S. W. Schmidt, P. Filippov, A. Kersch, M. K. Beyer, H. Clausen-Schaumann, *ACS Nano* **2012**, *6*, 1314–1321.
- [15] a) A. P. Wiita, S. R. K. Ainarapu, H. H. Huang, J. M. Fernández, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 7222–7227; b) S. R. Koti Ainarapu, A. P. Wiita, L. Dougan, E. Uggerud, J. M. Fernández, *J. Am. Chem. Soc.* **2008**, *130*, 6479–6487.
- [16] C. Börger, H.-J. Knölker, *Synlett* **2008**, 1698–1702.
- [17] M. Rief, F. Oesterhelt, B. Heymann, H. E. Gaub, *Science* **1997**, *275*, 1295–1297.
- [18] a) F. Oesterhelt, M. Rief, H. E. Gaub, *New J. Phys.* **1999**, *1*, 6; b) Y. Xue, X. Li, H. Li, W. Zhang, *Nat. Commun.* **2014**, *5*, 4348.
- [19] a) S. W. Schmidt, T. Christ, C. Glockner, M. K. Beyer, H. Clausen-Schaumann, *Langmuir* **2010**, *26*, 15333–15338; b) M. F. Pill, S. W. Schmidt, M. K. Beyer, H. Clausen-Schaumann, A. Kersch, *J. Chem. Phys.* **2014**, *140*, 044321; c) S. W. Schmidt, M. F. Pill, A. Kersch, H. Clausen-Schaumann, M. K. Beyer, *Faraday Discuss.* **2014**, *170*, 357–367.
- [20] M. K. Beyer, *J. Chem. Phys.* **2000**, *112*, 7307–7312.
- [21] a) L. Garnier, B. Gauthier-Manuel, E. W. van der Vegte, J. Snijders, G. Hadziioannou, *J. Chem. Phys.* **2000**, *113*, 2497; b) M. K. Beyer, *Angew. Chem. Int. Ed.* **2003**, *42*, 4913–4915; *Angew. Chem.* **2003**, *115*, 5062–5064; c) M. F. Iozzi, T. Helgaker, E. Uggerud, *Mol. Phys.* **2009**, *107*, 2537–2546; d) M. F. Iozzi, T. Helgaker, E. Uggerud, *J. Phys. Chem. A* **2011**, *115*, 2308–2315; e) J. Ribas-Arino, M. Shiga, D. Marx, *Angew. Chem. Int. Ed.* **2009**, *48*, 4190–4193; *Angew. Chem.* **2009**, *121*, 4254–4257; f) K. Wolinski, J. Baker, *Mol. Phys.* **2009**, *107*, 2403–2417; g) M. T. Ong, J. Leiding, H. Tao, A. M. Virshup, T. J. Martínez, *J. Am. Chem. Soc.* **2009**, *131*, 6377–6379; h) A. Bailey, N. J. Mosey, *J. Chem. Phys.* **2012**, *136*, 044102; i) U. F. Röhrig, I. Frank, *J. Chem. Phys.* **2001**, *115*, 8670–8674; j) D. Aktah, I. Frank, *J. Am. Chem. Soc.* **2002**, *124*, 3402–3406.
- [22] a) J. Ribas-Arino, M. Shiga, D. Marx, *J. Am. Chem. Soc.* **2010**, *132*, 10609–10614; b) H. M. Klukovich, T. B. Kouznetsova, Z. S. Kean, J. M. Lenhardt, S. L. Craig, *Nat. Chem.* **2012**, *5*, 110–114.

3.3. Additional Information

3.3.1. Elongations of Strained Bicyclic Mechanophores

The following section will deal with the approach that was taken to calculate the structural elongation of the macrocyclic 1,2,3-triazole mechanophore. In the results large deviations of up to 3.5 Å (25 %) between experimental values and theoretical predictions were found for the elongations. Such large errors are untypical for double zeta DFT treatments of molecular geometries. Furthermore, comparable deviations did not occur in comparable calculation by *Pill* and coworkers for the cycloreversion of a cyclobutane system^[81]. For both reasons possible causes for the discrepancies between experiment and theory are discussed in depths after the introduction.

The structural elongation ΔL of the macrocycle (see figure 2 and 3 in the publication) was evaluated by using COGEF potentials for the two model systems shown in figure 3.1. The fixed anchoring atoms for the COGEF protocol were the carbon atoms of the methyl residues. There are subtle differences between the model system and the mechanophore that was used in the SMFS experiments. Since the elongation ΔL solely depends on the lengths of the strained 1,2,3-triazole moiety and of the unfolded alcylic safety line, atoms that do not contribute to the length were omitted to reduce the computational cost. The omitted groups are the dangling alkyne- and azide residues for the unfolded system as well as the PEG spacer chains.

The fact that the dangling residues may be safely omitted in the calculations is closely related to a statement made above. The reaction path from the 1,2,3-triazole educt to the unfolded product state does not influence the length of the final structure. It is therefore impossible to discriminate between different reaction pathways from the experimental results.

Although the long flexible backbone of PEG chains will be elongated by a significant margin under mechanical loading forces, the polymere residues were ignored in the calculations. Since the structural lengths at the first and second fission event were evaluated at constant loading forces it was assumed that the tensile stress on the PEG linking the macrocycle to surface and AFM tip has no effect on the total elongation measured. This is because at equal forces the polymer elongation is assumed to be a constant and thus cancels out when taking differences between two structural lengths.

During the introduction to the theoretical foundations of CMC it was already discussed how external forces influence potential energy surfaces and distort molecular equilibrium geometries. Consequentially the elongation of a mechanophore must not be calculated

3. Triazole Mechanochemistry

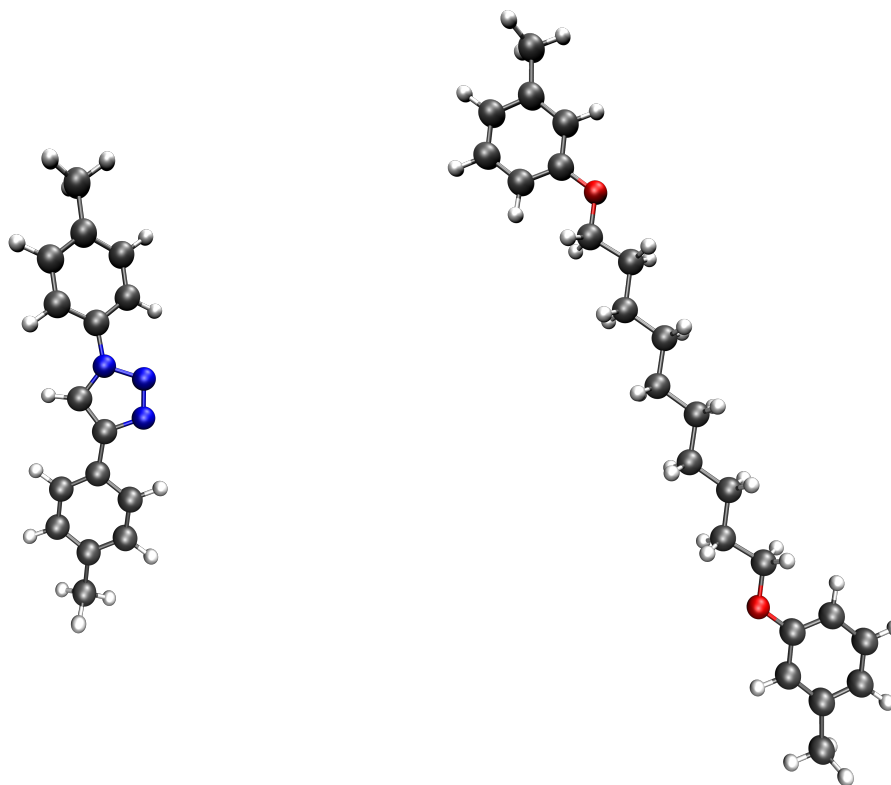


Figure 3.1.: Two model systems used to evaluate the change in length of the macrocycle in case of a fission event in the 1,2,3-triazole moiety. In COGEF calculations the frozen coordinate was the distance between the carbon atoms of the terminal methyl groups. In EFEI setups the additional gradient was applied to the same atoms. All unnecessary atoms are omitted.

by means of force-free equilibrium geometries but include the force explicitly. The external force on the the 1,2,3-triazole system was accounted for by calculating the COGEF potential for both model systems on the B3LYP/aug-cc-pVDZ level of theory over a wide range of distances between the anchoring atoms. The first derivatives of the COGEF potentials were fitted with a linear function to obtain a direct relation between the structural length of the model systems and the applied force. The difference of these force-dependent lengths, which is the desired quantity ΔL is plotted in figure 3.2. The experimental values obtained by SMFS experiments are plotted for comparison.

The B3LYP/aug-cc-pVDZ COGEF model has been validated with bigger basis sets and MP2 calculation on the smaller of the two model systems. The calculation sampled the potential in a region near the force-free equilibrium, single-reference methods were therefore assumed to be sufficient. Since the computationally more demanding

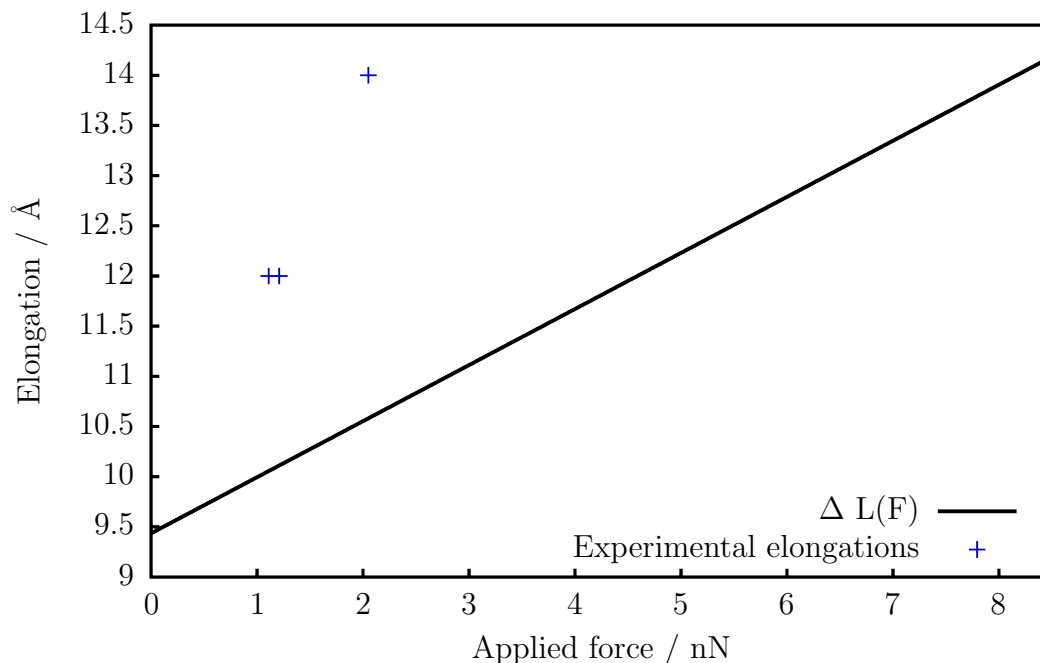


Figure 3.2.: Structural elongation obtained from linear fittings of the negative gradient of the COGEF potentials (black) and experimental values from reference^[2] (blue). It should be pointed out that no data points for the theoretical values are shown here, since the plotted function is the difference between two first-order fits of the structural lengths of the model systems. Data points for this plot can not be obtained because the force is an observable and not a scan parameter. The different stiffness of both COGEF potentials prevents defining a useful abscissa. The fitted original data can be found in the supporting information of the publication^[2] but is of no concern for the further discussion.

methods didn't yield qualitative changes in the investigated region of the PES, the B3LYP/aug-cc-pVDZ method was accepted as the best compromise between accuracy and computational expense.

The final first-order equations for the force-dependent lengths of the model systems obtained by the procedure outlined above are given in equation 3.1.

$$\begin{aligned}
 L_{\text{tria}}(F) &= F \cdot 0.1764 \frac{\text{Å}}{\text{nN}} + 10.5414 \text{ Å} \\
 L_{\text{line}}(F) &= F \cdot 0.6524 \frac{\text{Å}}{\text{nN}} + 20.0626 \text{ Å}
 \end{aligned}
 \tag{3.1}$$

As expected the force-dependent length of the 1,2,3-triazole L_{tria} has a smaller slope

3. Triazole Mechanochemistry

than that of the safety line. The smaller number of bonds between both anchoring points and the much more rigid five-membered triazole ring itself make the potential stiffer than that of the alicyclic chain. Therefore the force-dependent total elongation (eq. 3.2), which is the difference $L_{\text{line}} - L_{\text{tria}}$ retains a significant force-dependence.

$$\Delta L(F) = F \cdot 0.4760 \frac{\text{\AA}}{\text{nN}} + 9.5211 \text{\AA} \quad (3.2)$$

The results for the elongation conclude the introductory part of this section, the results and problems arising with them will now be discussed. As stated above there are large deviations between the experimental and theoretical results for the elongation of the system. The resulting elongation in equation 3.2 underestimates the experimentally obtained values significantly. The deviations range from 1.9 Å to 3.5 Å, which is up to 25 % of the total structural elongation and 17 % of the total length of the safety line structure. The linear plot in figure 3.2 shows that the lowest measured elongation of 12 Å is reached at an external force of 5nN which is already 2.5 times higher than the highest bond-breaking force that was measured.

DFT calculations in general, and DFT based COGEF simulations in particular, as shown in a recent publication by *Pill* and coworkers^[81], are capable of predicting structural elongations ΔL to a much higher accuracy. Therefore the remainder of this section is dedicated to the discussion of the origin of the artifacts mentioned. To anticipate the results: Although some sources of possible errors have been identified, the experimental results have not been successfully reproduced.

The Harmonic Approximation

When a linear relation between the exerted force F_0 and the structural elongation ΔL is employed, it is implicitly assumed that the underlying potential is harmonic and may be approximated by a second-order polynomial. Only in a harmonic potential the first derivative, which yields the force, is linear.

As the experimental results had bond-breaking forces in the range of 1-2 nN where static DFT estimates of the bond breaking force yielded 6-8.5 nN, the potentials were assumed to be only slightly tilted. In this region of small distortions potential energy curves can usually be well approximated with a second-order polynomial term.

On the other hand it may also be argued that the COGEF potential for both model systems is very soft near the equilibrium geometry. Stretching the geometry over a relatively large range yields only a slight increase in potential energy. Consequentially

3. Triazole Mechanochemistry

the gradient is small and the geometries can be stretched significantly with low external forces applied, maybe even far enough for anharmonic effects to become important.

The problem was addressed by using a third-order fit to the COGEF potential that captures anharmonic effects. The third-order term in equation 3.3 was fitted to the original COGEF potential used before.

$$V(x) = -\frac{1}{3}ax^3 + \frac{1}{2}bx^2 - cx + V_0 - F_0x \quad (3.3)$$

The first derivative yields the force acting of the anchoring atoms (eq. 3.4) and subsequently the force-dependent equilibrium distance (eq. 3.5).

$$F(x) = -\frac{dV}{dx} = ax^2 - bx + c + F_0 \quad (3.4)$$

$$x_0(F) = \frac{b}{2a} - \sqrt{\frac{b^2}{4a^2} - \frac{c + F_0}{a}}. \quad (3.5)$$

The calculated equilibrium distance x_0 is the length of the strained system, the desired elongation ΔL is again the difference of the force-dependent lengths of both model systems. The resulting force-dependent lengths that are obtained by fitting third-order polynomials to the B3LYP/6-31G* COGEF potentials used before are given in equation 3.6.

$$\begin{aligned} L_{\text{tria}}(F) &= 16.307 \text{ \AA} - \sqrt{6.930 \text{ \AA}^2 - F \cdot 0.8796 \frac{\text{\AA}^2}{\text{nN}}} \\ L_{\text{line}}(F) &= 29.401 \text{ \AA} - \sqrt{33.842 \text{ \AA}^2 - F \cdot 4.9279 \frac{\text{\AA}^2}{\text{nN}}} \end{aligned} \quad (3.6)$$

The result for the total elongation of the structure upon rupture of the 1,2,3-triazole moiety, is given in equation 3.7.

$$\Delta L(F) = 13.094 \text{ \AA} + \sqrt{33.842 \text{ \AA}^2 - f \cdot 4.9279 \frac{\text{\AA}^2}{\text{nN}}} - \sqrt{6.930 \text{ \AA}^2 - F \cdot 0.8796 \frac{\text{\AA}^2}{\text{nN}}} \quad (3.7)$$

The elongation term in 3.7 depends now on the force via a square-root term. This term arises from the fact that the force-dependent equilibrium lengths $L(F)$ are inverse functions of the gradient of $V(x)$ with respect to F , which is quadratic in x . The

3. Triazole Mechanochemistry

square root implies two things. First, it is undefined beyond the breaking force of the molecule, hence the plot in figure 3.3 ends at the breaking force 6.9 nN of the safety line, which is the less mechanically resilient of both model systems. Second, the slope of the elongation curve is larger and increasing with the applied force. This additional increase is the correction that is due to the third-order correction.

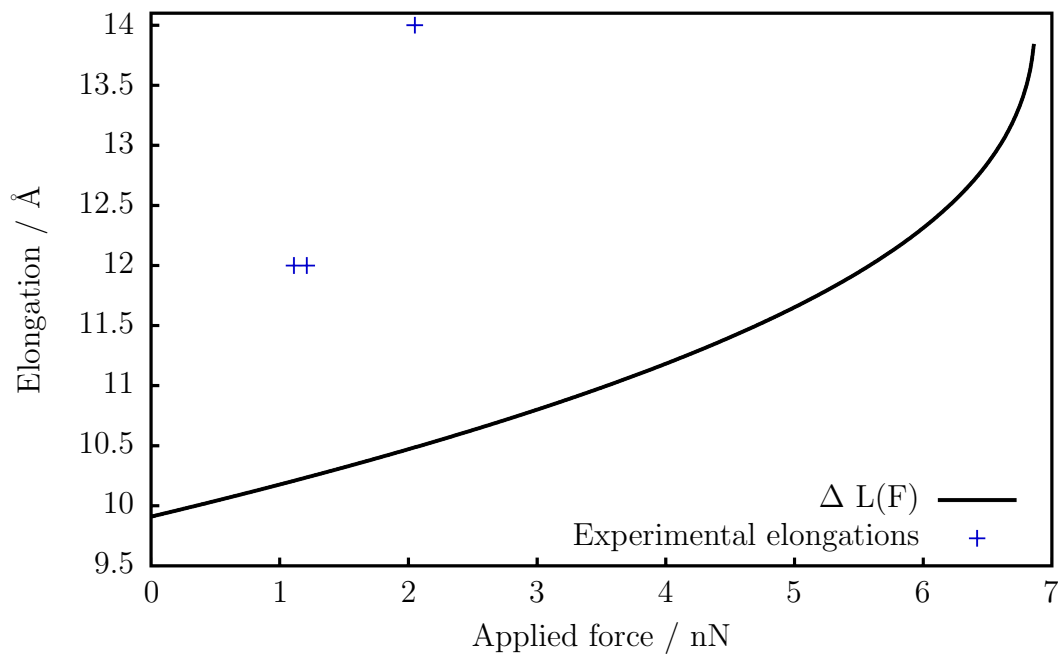


Figure 3.3.: Structural elongations obtained from the gradient of a third-order fit to the COGEF potential. The function is not defined at forces greater than 6.9 nN, as this is the estimated breaking force of the safety line.

Obviously the effect of this correction is negligible for the force region sampled by the experiment, the discrepancy between measured values and theoretical model remain substantial as table confirms.

Table 3.1.: Elongations of the structure calculated with the results from the third order fit. The forces used are the experimentally observed breaking forces.

Applied Force / nN	$L_{\text{tria}}(F) / \text{Å}$	$L_{\text{line}}(F) / \text{Å}$	$\Delta L(F) / \text{Å}$
1.11	24.07	13.87	10.21
1.21	24.12	13.89	10.24
2.05	24.53	14.04	10.49

3. Triazole Mechanochemistry

To summarize the previous section, it can be stated that the deviations between experiment and theory do not originate from neglecting anharmonic contributions to the COGEF potential. Even in the limit of the static breaking force of 6.9 nN the elongation is smaller than the experimentally obtained one. A less pessimistic view on the results above is that they show that the harmonic approximation for the COGEF potential may indeed be used without introducing significant additional errors to the elongation. This is because, for reasons discussed in the theoretical introduction to CMC, the experimental breaking forces are significantly lower than the static ones. This implies that if elongations are to be calculated for other systems in the future by using the isometrical approach, only a few points of the potential are needed since the fit is just linear. This opens up a wider range of quantum chemical methods like MP2, CASSCF or even Coupled Cluster to be used on the problem.

Isotensional versus Isometrical Approaches

The elongations calculated for the publication were obtained by using the isometrical COGEF approach. A number of points along the COGEF potential was calculated and the potential was fitted to obtain the forces and respective geometries. An easier and more straightforward approach to optimize strained geometries at a certain applied force is the EFEI method^[29]. EFEI was used by *Pill* to calculate elongations for a cyclobutene based system^[81].

In their publication *Pill* and coworkers investigated a bridged mechanoresponsive molecule of a similar type as the 1,2,3-triazole. In contrast to the results presented here, the elongations calculated by *Pill* show no systematic error towards smaller lengths, furthermore the highest relative deviation in their work is 5 % with an average error of 2.5 %. These are significantly smaller errors than those obtained for the 1,2,3-triazole system.

As mentioned above, one key difference between the calculations for the 1,2,3-triazole and *Pill's* calculations on the cyclobutane was the use of EFEI. In addition to that another residue was used to model the anchoring groups which mimics the system used in the actual experiments more closely.

Both of these approaches were used here to test whether they are able to improve on the 1,2,3-triazole results or not. Two independent sets of calculation were carried out. In the first set, EFEI was applied to the model systems shown in figure 3.1. In the second set of EFEI calculations, the methyl anchoring groups were substituted for amide groups to model the anchoring situation as found in the experiment more accurately.

3. Triazole Mechanochemistry

All EFEI calculations have been carried out with the TURBOMOLE^[30] quantum chemistry package, which was combined with a custom PERL script to add the external force vector before each geometry optimization step was started.

The first set of calculations, which was done on the unchanged model systems, consisted of twelve individual EFEI optimizations in total. The geometries were optimized for three experimentally obtained straining forces of 1.11 nN, 1.21 nN and 2.05 nN for the 1,2,3-triazole and the safety line system each (fig. 3.1). To estimate the effect of the basis set, the calculation were carried out with the STO-3G minimal basis set and *Ahlrich's* def2-TZVP^[82-85] using the B3LYP density functional and TURBOMOLE's RI-DFT^[86-89] implementation. The basis set was changed from *Dunning's* generally contracted correlation consistent basis set which was used in the GAUSSIAN09 calculations before, because TURBOMOLE works more efficient with segmentally contracted basis sets. Furthermore *Ahlrich's* auxiliary basis sets are recommended to be used with def2 basis sets in RI-calculations.

In the second set the amide groups were added to model the anchoring situation more accurately. Again twelve optimizations were done. This time the density functional was varied and the basis set was kept unchanged. One set of optimizations was done with the RI-B3LYP/def2-SVP method while the second set was calculated on the RI-M06-2X/def2-SVP level of theory. The def2-SVP basis set was chosen because *Pill* used the B3LYP/def2-SVP method in his calculations.

Table 3.2.: Structural elongations calculated in the EFEI formalism using different DFT methods and model systems.

Methyl anchor	B3LYP/STO-3G	B3LYP/def2-TZVP
Applied Force nN	$\Delta L(F) / \text{\AA}$	$\Delta L(F) / \text{\AA}$
1.11	10.38	10.23
1.21	10.40	10.26
2.05	10.61	10.48
Amide anchor	B3LYP/def2-SVP	M06-2X/def2-SVP
Applied Force / nN	$\Delta L(F) / \text{\AA}$	$\Delta L(F) / \text{\AA}$
1.11	10.76	10.67
1.21	10.78	10.76
2.05	10.94	10.87

The results in table 3.2 indicate that whether the isometrical or the isotensional approach is used the results remain almost unchanged. Due to the *Legendre* transform relation discussed before that was to be expected. Furthermore it has been shown that

3. Triazole Mechanochemistry

the basis set and the functional used has no influence on the results. The most significant change was due to the introduction of the amide anchoring groups. However, a quite significant maximal deviation of 3.1 Å or 22 %, when compared to the experimental elongation of 14 Å, remains.

Dynamical Effects

In the publication^[2] it was argued that the remaining systematic error can be attributed to thermal expansion in the anharmonic potential or dynamic pooling of energy in the length-determining degrees of freedom, which can not be accounted for by static calculations at a temperature of 0 K in vacuo.

The later publication by *Pill* makes these arguments invalid, since the systematic error does not occur for an almost identical experimental setup with a theoretical treatment at an equal level of accuracy. If anything the longer safety lines used there should lead to an even larger systematic error.

Also it was already shown above that anharmonic contributions in the investigated force regime have only small effects on the elongation. Larger contributions could only have been expected from softer degrees of freedom like valence angles and torsional angles. However, it was convincingly argued by *Marx* and coworkers that any conformational rearrangements happen at forces at least one order of magnitude lower than the force needed for covalent processes to occur^[90]. This means all possible conformational changes are completed as soon as straining forces in the nanonewton-regime are reached. Valence angles along the force-loaded molecular backbone are strained to the point where their potential stiffness equals that of the strained bonding potentials scaled by an angular factor that accounts for the direction of the force vector. Therefore anharmonic contribution to the elongations from these strained angles can be expected to be very small, in the same range as seen for the anharmonic bond contributions, by far too small to account for the large deviations in the range of 3 Å.

Exemplary calculations for the dynamic elongations with methods of molecular dynamics were done with the reactive force field REAXFF as implemented in the LAMMPS^[31] software suite with the parameterization by *Mattsson* and coworkers. The results were inconclusive and inconsistent, which was attributed to the fact that the parameterization was not optimized to handle strained molecules or 1,2,3-triazoles and visibly struggles with the latter.

As no improvement to the results was expected by molecular dynamics simulations no effort has been made to set up computationally expensive DFT-based ab-initio MD.

Conclusion

Different theoretical models and approaches have been used to try and improve on the the quality of the results obtained with the harmonic COGEF approach. None of the approaches above succeeded in reproducing the experimental results more accurately.

The consistency of the theoretical results through various methods, basis sets and simulations approaches however is somewhat remarkable. All used methods converge to the same results and the calculations are operating in a region of the PES where single-reference methods are reliable. Therefore the calculated elongation for the model systems presumably can be trusted.

This leaves two possible conclusions. First, the experimental procedure may have introduced an error that overestimates the elongation by 2-4 Å. Second, and far more interesting, the theoretical underestimation of the elongation stems from an effect intrinsic to the 1,2,3-triazole macrocycle not yet considered.

3.3.2. Mechanical Considerations Regarding Mechanically Facilitated Retro Click Reactions

Huisgen's 1,3-dipolar cycloaddition^[80] or AAC is a pericyclic reaction that couples an azide and an alkyne to form a five-membered 1,2,3-triazole moiety. Depending on the regioselectivity of the reaction, the product can be a mixture of 1,4- and 1,5-disubstituted triazoles. Based on the publication by *Bielawski* and coworkers it was assumed that the AAC can be reversed by applying mechanical strain to the 1,4-disubstituted 1,2,3-triazole moiety. Early COGEF calculation raised doubts if an reversed AAC would occur when the 1,2,3-triazole mechanophore is mechanically strained.

The COGEF potential shown in figure 3.4 suggests that mechanical stress rather leads to a dissociation of the covalent bond connecting the nitrogen atom of the triazole ring with the anchoring groups than inducing a cycloreversion reaction. These early doubts were at the time dismissed as a probable artifact of the simplistic unrestricted B3LYP/6-31++G** model treatment, which is known to be unreliable for dissociations due to its single-reference nature.

The suspicion that there might be something wrong with the mechanism of the reversed AAC was then supported by the retraction of the original publication by *Bielawski* and coworkers due to scientific misconduct and the scarcity of positive experimental results for the 1,2,3-triazole. To this day it is therefore unclear whether the AAC was reversed due to strain in the AFM in the three positive experimental double rupture

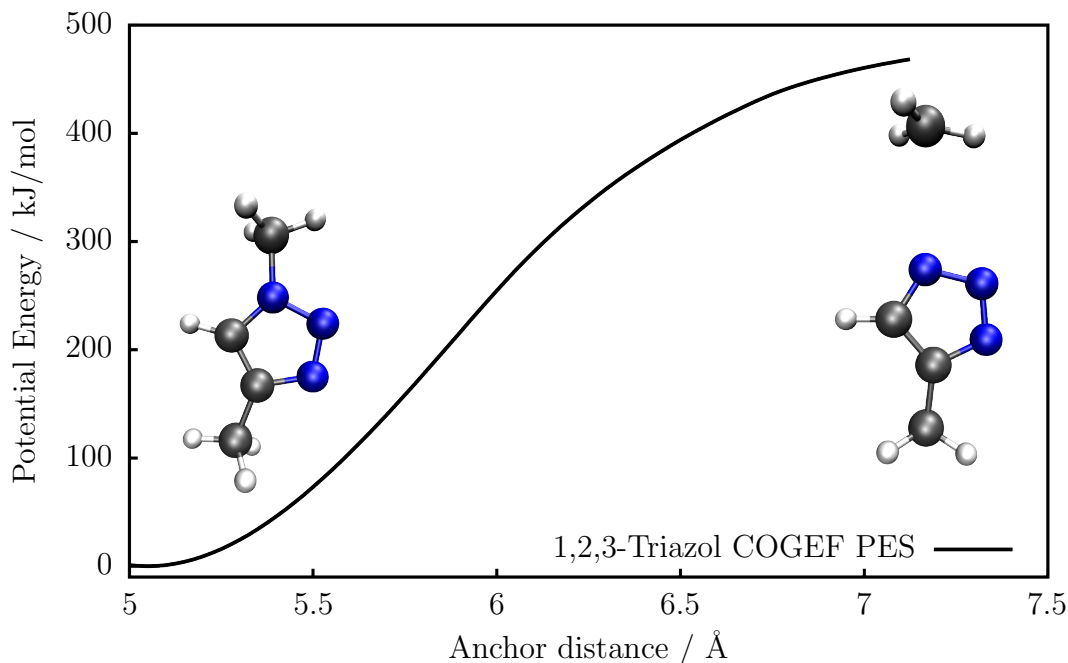


Figure 3.4.: COGEF potential energy curve of dimethyl substituted 1,2,3-triazole using the methyl carbon atoms as anchoring atoms.

events or if a homolytic fission at the 1- or the 4-position of the triazole took place.

Quite recent calculations by *Jacobs* and coworkers show that the reversed AAC indeed competes with the homolytic fission of the anchoring groups^[91]. In case of the bridged 1,2,3-triazole considered here however, the results of *Jacobs* do not apply. The reaction investigated there concerns another type of AAC reaction, the so called strain-promoted AAC or SPAAC^[92,93], where the geometric strain of a cyclooctine ring is used to lower the activation barrier for the AAC to occur. The experiments within project A05 utilize terminal azide and alkyne residues for the click-reaction. Furthermore *Jacobs* compares 1,4- and 1,5-substituted triazoles while the macrocycles used by *Schütze* are exclusively 1,4-substituted. Finally the theoretical methods used by *Jacobs* and coworkers, namely B3LYP/6-31G and B3LYP/6-31++G, were not validated for the dissociation reaction of the azide cyclooctine adducts and may, for their single-referential nature and unpolarized basis sets, lack the necessary accuracy for a quantitative investigation of competing reaction paths.

Consequently it was deemed necessary to investigate the cycloreversion mechanism in the scope of the SMFS experiments done by *Schütze*.

The investigation was started with the search for a transition-state geometry, since an internal reaction coordinate energy curve (IRC) connecting the 1,2,3-triazole state

3. Triazole Mechanochemistry

with the isolated azide-alkyne educt minimum promised to grant deeper insight into the reaction. The transition state was optimized using the QST3 method^[94] as implemented in GAUSSIAN09^[95]. Since the AAC is a pericyclic reaction, the transition state was assumed to be a cyclic complex of the alkyne and azide educts. The calculation converged reasonably and the normal coordinate analysis showed the expected imaginary mode. The structures and estimated energetics for the reaction on the unrestricted B3LYP/6-31++G** level of theory are depicted in figure 3.5.

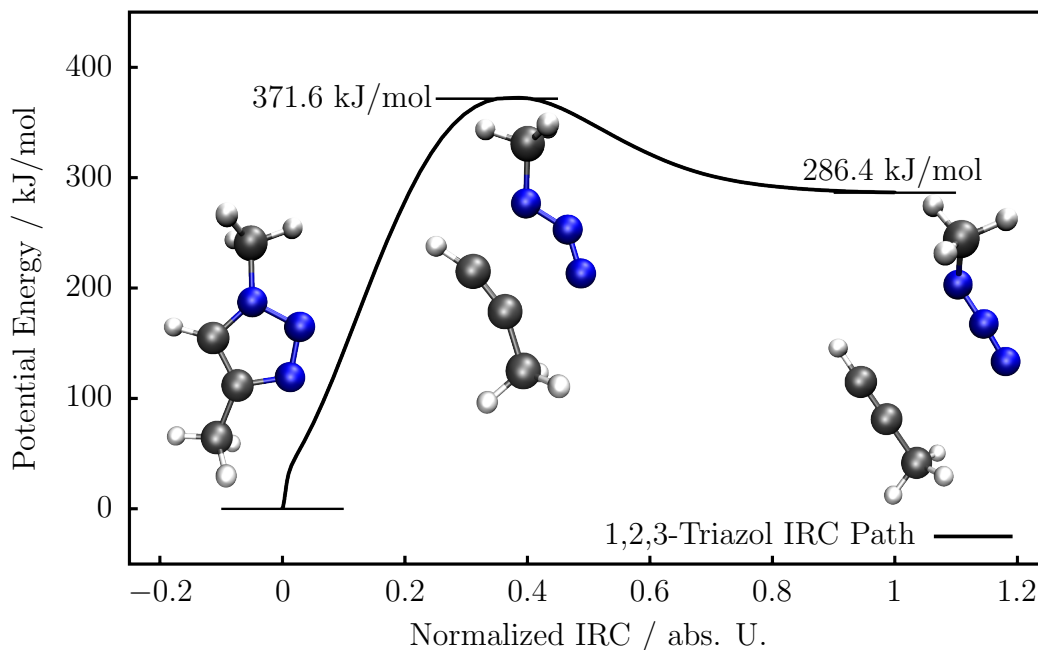


Figure 3.5.: Energetics along the internal reaction coordinate for the cycloreversion of a 1,4-methylized 1,2,3-triazole.

Before discussing the results, a note on the reliability of the B3LYP/6-31++G** is in order. Since the simulation encounters regions of the PES with significant multi-reference character due to the formation of covalent bonds, the single-reference DFT treatment can not be trusted. To validate the method the curve shown in 3.5 has been calculated with various wave function methods and basis sets, including a full-valence CASPT2/aug-cc-pVTZ calculation. While the relative energies of educts, transition state and products varied by roughly 10 kJ/mol, the overall shape of the PES remained unchanged. It was therefore concluded that a double-zeta DFT treatment is sufficient for a qualitatively correct modeling of the reaction.

The normalized reaction coordinate shown on the abscissa in 3.5 requires an explanation. In contrast to the COGEF potential in figure 3.4 a one-dimensional reaction

3. Triazole Mechanochemistry

coordinate is not easily defined for the IRC path. The reversed AAC occurs via a concerted lengthening and ultimately cleavage of both CN-bonds between the 1- and 3-nitrogen and the respective neighbouring carbon atoms in the ring. The length of both bonds is approximately equal at different points of the IRC path, while the remaining structure shows almost no atomic movement. The one-dimensional reaction coordinate was therefore assumed to be the diagonal of a rectangle spanned by the bond elongations in one IRC step. This choice has two implications. First, the unit of the reaction coordinate has not the same meaning as for example the stretching along the COGEF potential curve. Hence the coordinate was normalized, where 0 represents the converged educt structure and 1 the product structure along the IRC path. Second, the concerted lengthening of the bonds is not dominant for all regions of the IRC path, in the beginning and the end energy is mainly used to deform the softer degrees of freedom. This makes the reaction coordinate a bad choice for the first few and the last few points of the IRC path. This can be seen in the beginning of the IRC where figure 3.5 shows a sharp increase in the energy for a small change in the reaction coordinate. This is an artifact due to the fact that the bonds are not lengthened in that region, but the 1,2,3-triazole ring is deformed. These points should consequentially be ignored in the discussion of the results. In practice this is no severe problem since the center part of the IRC, where the potential really is steep and the transition state is located, is of interest for the following discussion. In this region the reaction coordinate is good and allows for estimates of activation forces.

The potential curve in figure 3.5 shows an activation barrier of 370 kJ/mol to reach the transition state. With an estimated dissociation energy of 470 kJ/mol (fig. 3.4) the reversed AAC is energetically preferred by a margin of 100 kJ/mol. Since an external force has a profound impact on the shape of the PES this preference of the reaction path may change rapidly when the reaction is induced by means of force and not thermal activation. Various theoretical works have shown that products and reaction paths that are inaccessible by thermal activation open up when the system is subjected to mechanical force^[13,96,97].

An estimate of the activation force for the COGEF potential is obtained by its first derivative. The maximum of the first derivative shows a critical force of 6.49 nN for the homolytic fission of the anchor CN-bond.

For an estimate of the activation force needed to induce the reversed AAC, the reaction coordinate definition discussed above is used. The force obtained this way is shown in figure 3.6. The first points in this plot are omitted since the definition of the

3. Triazole Mechanochemistry

reaction coordinate would yield an excessive gradient for reasons discussed above. The activation force, which is the maximum where both bonds are lengthened to roughly 2 Å, is estimated to 7.37 nN.

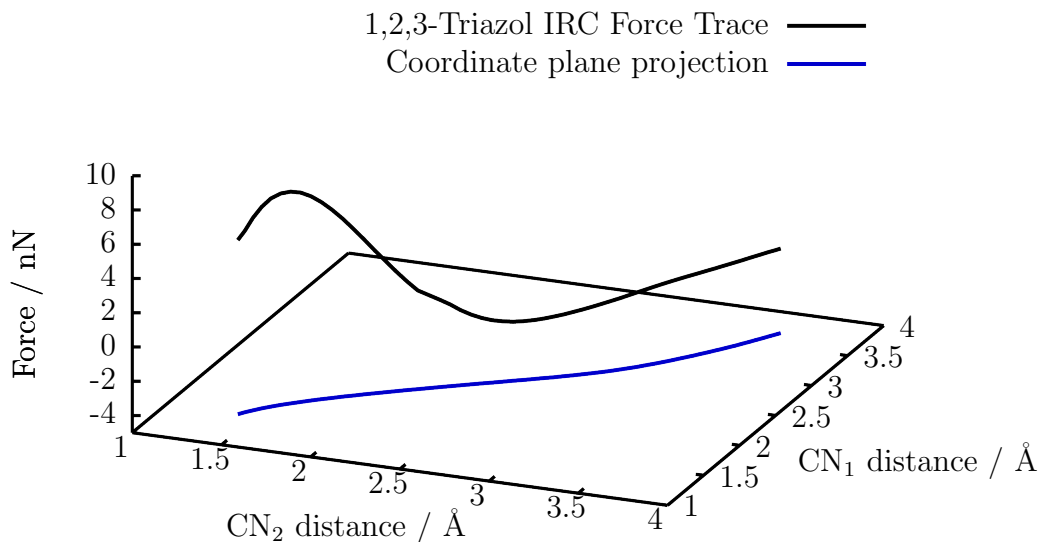


Figure 3.6.: Gradient trace, i. e. Force, traced out by the cycloreversion reaction in the space spanned by the two CN bond lengths in the triazole ring. To improve readability of this 3d plot, the projection on the coordinate plane was also plotted in blue.

These results suggest that the homolytic fission of the anchor is activated at lower external forces than the reversed AAC and therefore the prevalent mechanism in the SMFS setup. However, the single-reference wave function used to determine both activation forces may not be reliable enough for quantitative gradient results in the critical region. Furthermore the approximation of the reaction coordinate introduced to derive the force adds to the error in the estimation. Thus it would be dangerous to conclude the homolytic fission is preferred. For the time being, it is safe to state that both activation forces are within the same order of magnitude and that it can not be decided unambiguously with path is preferred.

A final conclusion can be reached by accounting for the fact that the applied force is a directed quantity. The work, or in this case energy, transferred to a degree of freedom is given as the dot product of an acting force vector and the direction vector of the strained coordinate. Therefore an effective force F_{eff} can be calculated for the model considered

3. Triazole Mechanochemistry

here. The effective force is the external force F_0 multiplied with the cosine of the angle between the force vector and the strained degrees of freedom.

In case of the COGEF potential that leads to the homolytic fission, the angle between the anchoring CN- and CC-bonds and the applied force is in the range of 5° . The cosine of 5° is almost 1 and therefore the total applied force is acting on the anchoring bonds in an SMFS experiment.

The geometries shown in figure 3.5 indicate that in case of the reversed AAC reaction the applied force F_0 is steeply angled against the important CN-bonds within the ring. This behaviour is schemtically shown in figure 3.7. The reaction coordinate defined before allows to calculate that angle, which is 50° on average throughout the reation path. The effective force F_{eff} acting along that reaction coordinate is now only a fraction of 64% of the applied force F_0 . In other words the applied force F_0 has to be 1.55 times the previously estimated activation force of 7.37nN to activate the reaction.

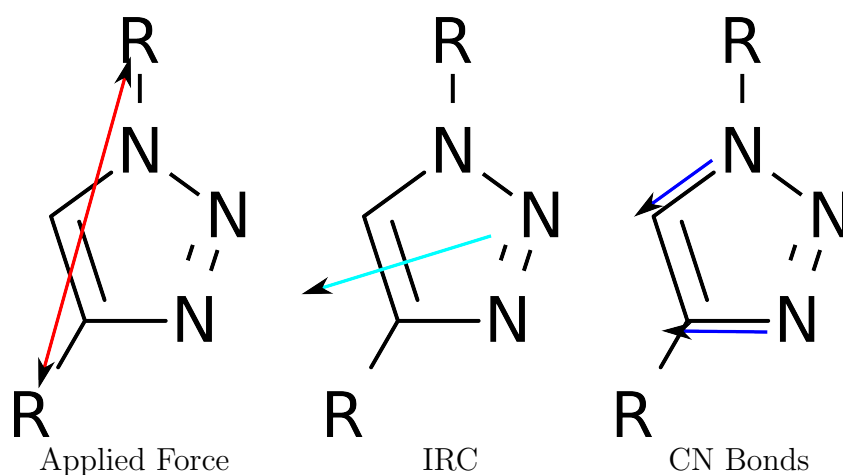


Figure 3.7.: Schematic representation of the step angling of the IRC against the applied mechanical force. The direction of the applied force is shown on the left in red. The lengthening direction of the reactive CN-bonds is indicated by blue arrows on the right and the sum of both vectors in cyan in the middle structure. The angle between the applied force and the reaction coordinate for the concerted opening of the CN-bond enclose an angle of 50° . This two-dimensional representation is reasonably accurate since the system is planar.

This final activation force of 11.5nN is considerably higher than the activation force for the homolytic fission, in fact so high that this reaction path can be considered to be impossible in the setup described here. Furthermore the activation force is higher

than the estimate for any breaking force of a single bond^[24]. This also implies that the reversed AAC for the 1,4-disubstituted 1,2,3-triazole can not be observed in any strained system containing single bonds along the strained coordinate.

Modification of the 1,2,3-triazole

To overcome the problem of high activation forces, Prof. Dr. Lünig suggested a modification of the 1,2,3-triazole unit. He suspected that the functionalization in 3-position with an additional methyl group could significantly lower the activation force of the reaction. To investigate the mechanism, the same approach as before was chosen. The QST3 algorithm was used to locate the transition state for the reaction, after that an IRC path was calculated.

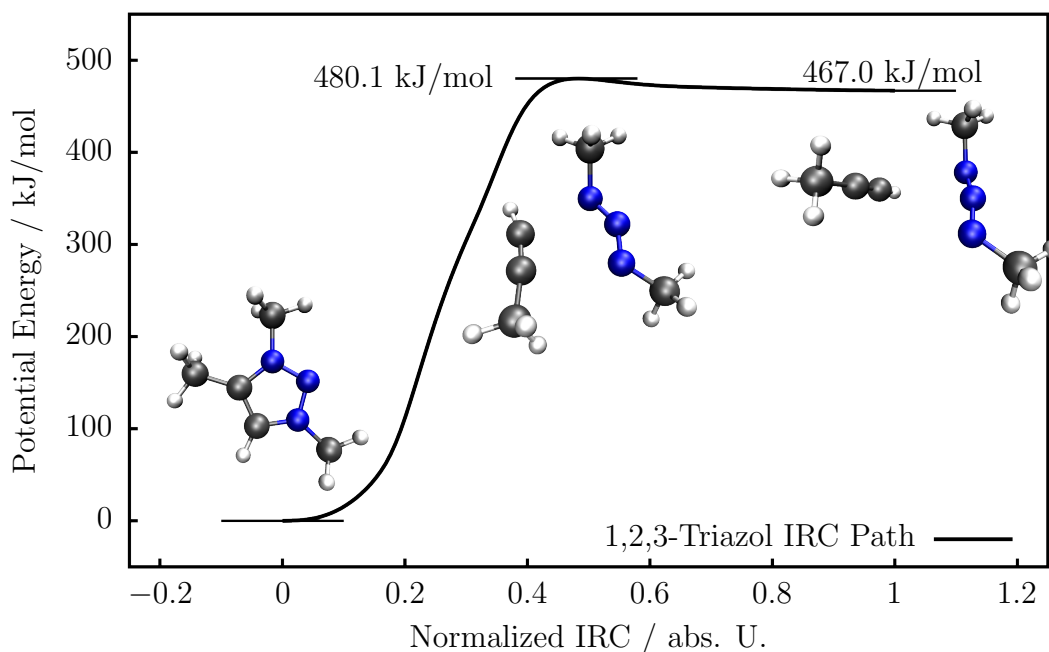


Figure 3.8.: Energy path along the IRC of the methylized 1,2,3-triazole. The activation barrier is higher when compared to the non-methylized triazole, and the local energy maximum at the transition state is less pronounced.

The results compiled in figure 3.8 are indicating that gradients and activation barriers remain almost unchanged when comparing them to the unfunctionalized system in figure 3.5. Therefore the behaviour of the modified system in an AFM experiment is expected to be unaltered. It was subsequently concluded that the modified 1,2,3-triazole would be no viable target for organic synthesis.

Conclusion

In conclusion to this section it may be stated that the design of the experiment had some hidden difficulties that made it hard to obtain satisfying results. The geometry of the mechanophore had multiple possible breaking points instead of just one and it was therefore impossible to decide ad hoc which breaking point was addressed in the experiment. Furthermore the transfer of the applied force into the 1,2,3-triazole unit at the core of the mechanophore was not ideal as the reaction pathway of the reversed AAC is steeply angled against the inclined force vector. Another issue concerns the activation of the cycloreversion. The activation force as well as the activation energy is comparable to the critical forces and dissociation energies of single bonds in the mechanophore and the polymere anchor backbone, in some cases even higher. Therefore the mechanophore can not be specifically addressed by an external force.

Three of the above mentioned problems have been addressed in the work by *Pill* and coworkers^[81]. The bridged cyclobutane mechanophore used there is constructed in such a way that double ruptures in the force extension curves can only occur if the mechanophore undergoes a cycloreversion (fig. 3.9). This way every successful experiment is guarantueed to correspond to a cycloreversion. The cyclobutane is a strained structure as the angles of the sp^2 -hybridized carbon atoms in the four-membered rings deviate significantly from their preferred tetrahedral angles. The activation force and energy is therefore greatly reduced compared to the very stable 1,2,3-triazole moiety. The last and maybe most important difference is that the anchoring groups that transfer the force in the mechanophore are attached adjacent to one another. Therefore the cycloreversion of the cyclobutane occurs via a two-step mechanism in a zipper like fashion. The reaction coordinate progresses over two bond dissociations where the force acts almost parallel to the bond. This is in contrast to the situation in the 1,2,3-triazole where two bonds are broken at the same time with the acting force steeply angled against the reaction coordinate.

The effect of this improvement shows in the results. The yield of nine positive experiments that could unambiguously assigned to cycloreversions proves that mechanophores can be addressed specifically to control chemical reactions on the molecular level by external mechanical forces.

As the activation force for the 1,2,3-triazol is too high to induce the reaction in an AFM experiment the system may be modified to reduce the necessary forces. This can be achieved by making use of the SPAAC discussed in *Jacobs* work^[91]. A strained 1,5-substituted 1,2,3-triazol like the cyclooctine adduct could yield much better results as

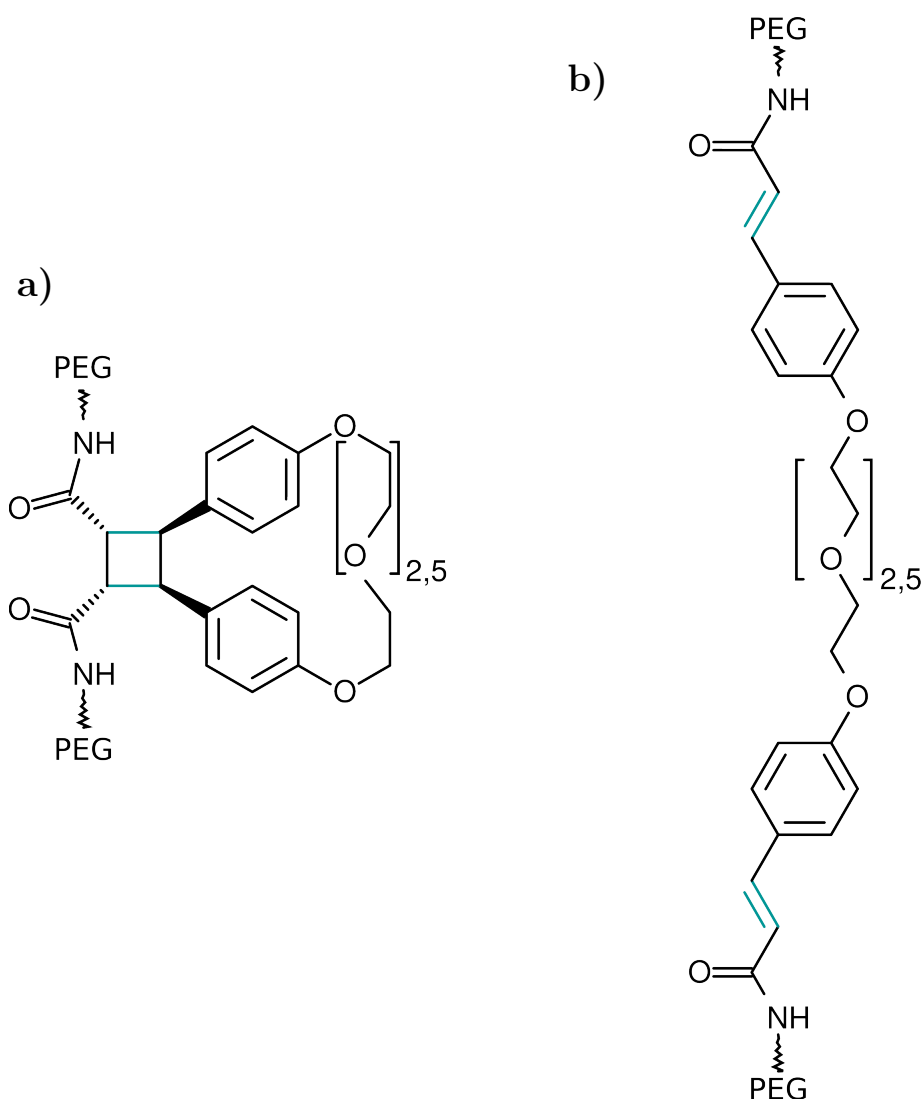


Figure 3.9.: Cyclobutane structure as used in the work by *Pill* and coworkers^[81]. The closed structure in panel a) is strained by an external force applied over the PEG spacer chains. When the activation force for the cycloreversion is reached the cyclobutane dissociates and the structure unfolds to the open form in panel b). The forming double bonds from the cyclobutane are indicated in cyan.

mechanophore in the macrocycle. The activation energy would be lower than in case of the terminal alkynes and the zipper-like geometry allows for a two-step rupture with the applied force acting parallel to the strained bond.

The experimental yield may also be increased by substituting the central moiety of the mechanophore. Dichalkogenides, which are discussed in a later chapter, like disulfides or diselenides show low activation forces and dissociation barriers. If incorporated in the

3. Triazole Mechanochemistry

macrocycle they should perform well in the SMFS experiments.

Another option would be to abandon the covalent concept of the experiment in favor of supramolecular adducts that couple via molecular recognition mechanisms. Hydrogen bonds, which are the central interactions in molecular recognition, are roughly one order of magnitude weaker than covalent bonds and therefore easily activated by mechanical forces. Such supramolecular mechanophores furthermore may be reversibly switched between the bound and the unbound state.

From the theoretical point of view it would be interesting to refine the model introduced in section 2.1 and confirm whether the experimental SMFS results for the 1,2,3-triazol can be predicted. If the ratios of successful experiments can be reproduced by applying a kinetical model it would be much easier to design promising machanophores in silico.

The model introduced by *Beyer*^[24] that is used to explain experimental findings today^[81,91] uses approximations that are not satisfied in some application cases, as in the present one. This makes estimated reaction rates and lifetimes problematic. In fact the method predicts lifetimes longer than the age of the universe for the 1,2,3-triazole mechanophore when strained with a force of 1.11nN, which was the lowest breaking force that was experimentally obtained.

The problem with the model is the underlying assumption that the COGEF potential is well approximated by a *Morse* potential. The critical force F_c as well as the dissociation energy D_e are calculated and the force constant β is then fitted according to the formula $F_c = \beta D_e / 2$. The resulting potential of course has the correct dissociation energy and maximal slope but may feature flawed curvatures. This problem leads to wrong estimates of the force-dependent reduction of the activation barrier $V_{\text{eff}}(F)$ that strongly influence the lifetimes, especially in the regime of low straining forces.

A better choice for a model potential for strained polymeres would be a chain of coupled *Morse* oscillators with different force constants per bond and maybe angular terms accounting for the bonding angles. In case of cyclic structures it would be even more difficult to find an appropriate model potential. The COGEF potential would then have to be evaluated at many points including the repulsive part to obtain a meaningful set of fitted parameters.

An alternative approach to this mathematically involved and computationally demanding procedure would be optimizing geometries for the equilibrium and transition state at different straining forces^[27] with the EFEI method and evaluate the lifetimes pointwise.

4. ReaxFF Parametrization and Disulfide Mechanochemistry

As was already stated in the previous chapter, the aim of subproject A05 within the SFB677 was to synthesize and investigate molecular switches that react upon mechanical stimuli. As an alternative design to the 1,2,3-triazole-based mechanophore discussed before, a disulfide system was proposed. The disulfide moiety, or more generally dichalkogenides, were chosen for several reasons. The disulfide group is well understood in the scope of protein chemistry and was shown to break easily under mechanical strain^[10]. Hence the disulfide bridge is a reasonable choice for designing molecular PBPs. Furthermore sulfides have been shown to conduct electrical currents in single-molecular junctions (SMJs)^[9]. Mechanoresponsive disulfides may therefore be the basis for materials that change their conductive properties upon mechanical strain to function as sensors. It is also possible to think of switchable SMJs although the mechanical stimulus would be difficult to exert to the mechanophore and photoswitchable SMJs are more promising for applications^[98]. Last but not least disulfides were deemed interesting because it was assumed that the synthesis of bicyclic systems analogous to those seen with the triazole moiety can be done by using well-known coupling reactions.

The theoretician's role in the project was to model the covalent mechanochemical conversions that occur in the AFM experiments. A dynamical description with the option for explicit treatment of different solvents was desired.

To this date the accurate description of mechanochemistry remains a challenging task in theoretical chemistry. There are several obstacles posed by mechanochemistry which are not present in many other cases. On the one hand mechanophores of interest are often big systems like proteins or polymeres in solution or adhered to complex surfaces, on the other hand the reactions often involve formation and breaking of covalent bonds. Due to the size of those systems they are not tractable by high-level ab initio methods or DFT on small- to medium-sized computing systems. The covalent character disqualifies them for treatment with traditional MM (i.e. nonreactive) methods. In fact the situation

4. ReaxFF Parametrization and Disulfide Mechanochemistry

is even worse, since the PES regions of interest for covalent mechanochemistry usually show significant multireference character, hence reliable results can only be expected when using multiconfigurational approaches.

Out of this rather uncomfortable perspective on theoretical mechanochemistry the motivation for introducing a reactive force field formalism backed by a powerful and robust tool for parametrization to this area of research was born. The problem would then be converted from the description of the mechanochemical phenomena with computationally expensive high-level methods to the calculation of a small subset of reliable reference data and fitting the force field parameters to them. With the preexisting experience on global optimization and REAXFF in the *Hartke* group as well as the in-house developed evolutionary algorithm package OGOLEM the author was in the ideal position to pursue this task.

However, fitting a REAXFF parameter set to a representative subset of data for the reactions of interest is not as trivial as it may seem. Using a reactive force field may circumvent most of the above-mentioned problems, but the choice of reference data is not at all clear. In the end it all comes down to a trial and error procedure to find out what data should be included in the reference set and which should not.

Furthermore the resulting objective function defined by parameters and reference set, as was already mentioned in the theoretical introduction, is extremely vast and has a very challenging structure. Another problem arises from the fact that REAXFF has a large number of adjustable parameters, therefore, due to the typically quite limited sizes of the reference sets, overfitting is an issue. This limited size especially is the case when high-quality MR data is used as reference set as the calculations have high computational demands even for small systems.

The difficulties mentioned above were tackled over the course of the project. The results have been published in two papers.

The first publication by *Dittner* and coworkers^[1] focusses on the optimization algorithm itself. The OGOLEM package was expanded to allow efficient treatment of evolutionary optimization for REAXFF parameters. To this end the preexisting OGOLEM software was interfaced with the sPUREMD code by *Aktulga*. With this combination a speedup of roughly one order of magnitude over an older implementation^[99,100] as well as more stable results were achieved.

The second publication by *Müller* deals with the construction of a force field for disulfide-based applications in more detail. Assembly of the reference set and adjustable parameters as well as overfitting are discussed. The resulting parameter set was then

4. *ReaxFF Parametrization and Disulfide Mechanochemistry*

used to conduct proof-of-principle calculations.

4.1. **Publication: Efficient Global Optimization of Reactive Force-Field Parameters.**^[1]

Contribution to the paper:

- Setup of a benchmark reference set.
- Setup and execution of benchmark calculations with the older GA/ADF software package.
- Setup and execution of benchmark calculations with the new OGOLEM/SPUREMD.
- Comparison of the capabilities of old and new setups.
- Optimizing the OGOLEM/SPUREMD input for best Results.
- Major contributions to the publication text.

Full text reprinted with permission. Copyright 2015 Wiley.

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 EA s 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

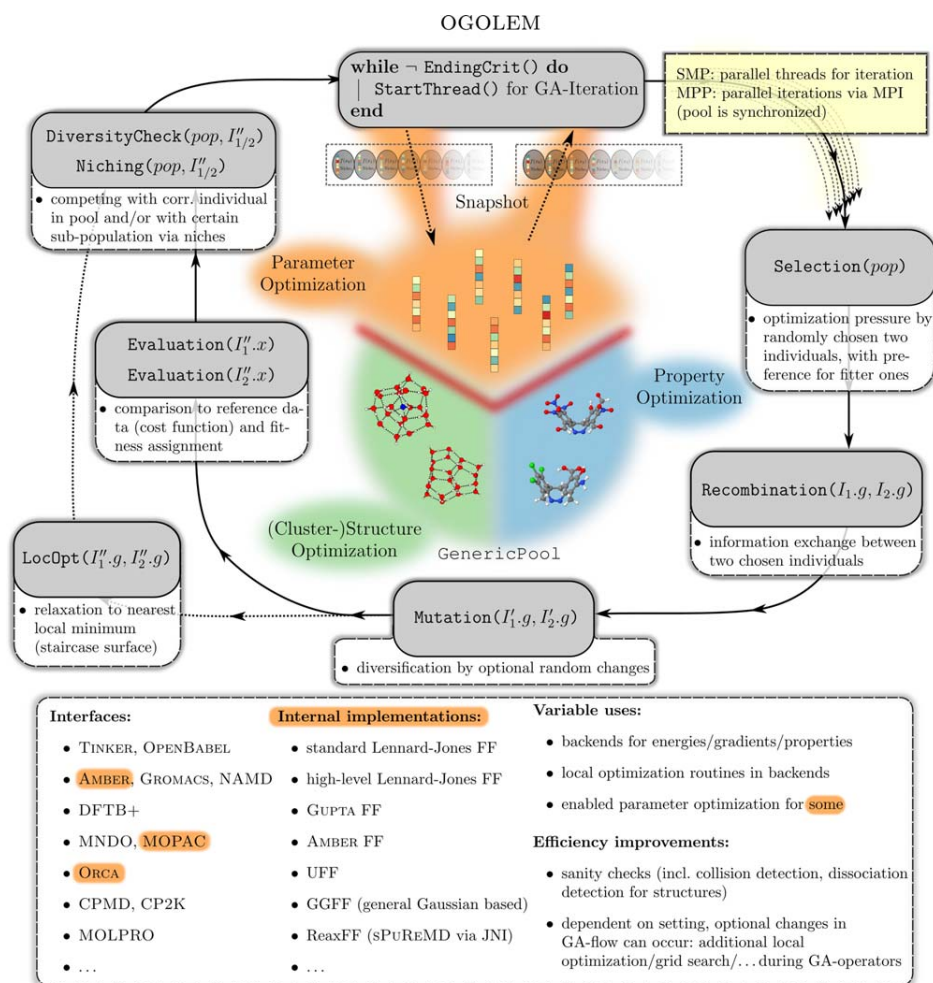


Figure 1. Flowchart of GA-iterations within OGOLEM (clock-wise). Certain important GA-operator steps together with their abstract role are shown. Many different implementations for these single operators are available, like described in the main text and also published in the cited literature of this 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.

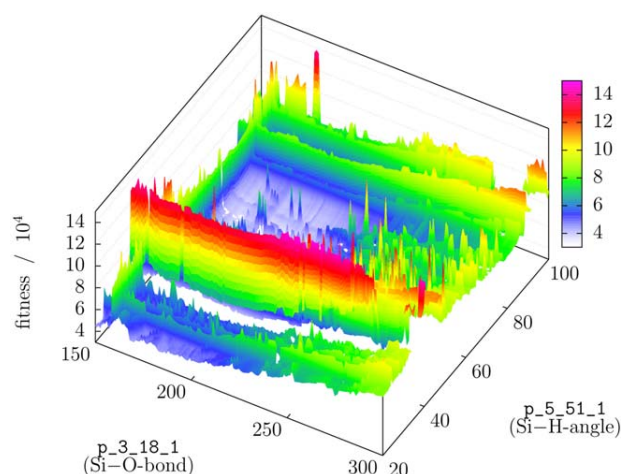


Figure 2. 2D objective function surface for two parameters (out of 67) of the SiOH training set. An intermediate solution with an error sum of about 100,000 is shown, occurring during a GA-run. Some interpolation due to smooth color progressions is implied. Transparent regions are erratic or mountain-like “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

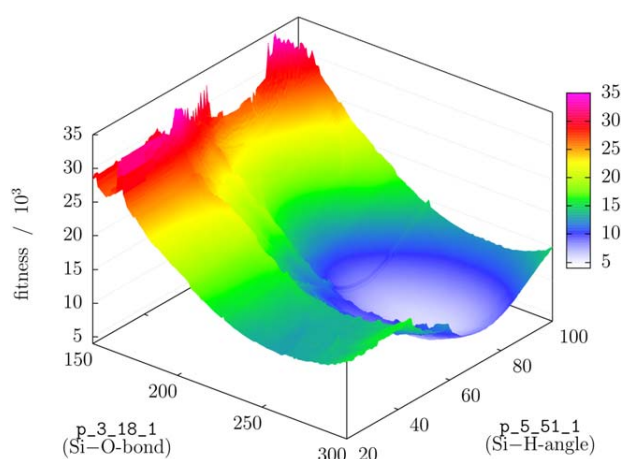


Figure 3. 2D objective function surface of the same two parameters as in Figure 2, but for a good solution near the end of a GA-run, with an error sum of 6150 (close to the global minimum). Again, some interpolation due to smooth color progressions is implied. [Color figure can be viewed in the online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

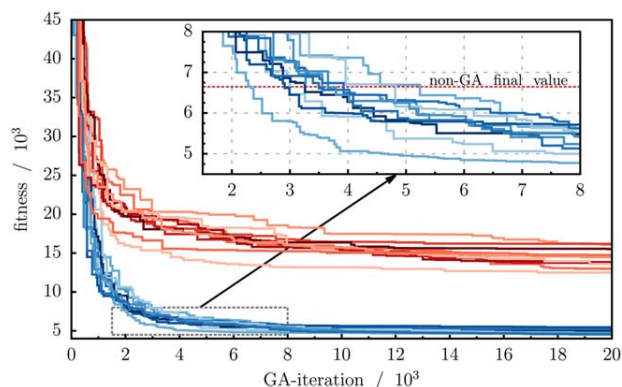


Figure 4. 10 selected but representative GA-runs with our old code (red lines)^[34,35] for the SiOH benchmark, in comparison to 10 with our new OGOLEM/sPuReMD-combo (blue lines, present work). The objective function value (i.e., error sum that is used as fitness for our GA) for the best individual of 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 Opteron 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]}	<i>run1</i> ^{[c],[b]}	<i>run2</i> ^{[d],[b]}	<i>run3</i> ^{[e],[b]}	<i>run4</i> ^{[b],[f]}	<i>run5</i> ^{[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 Opteron 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.

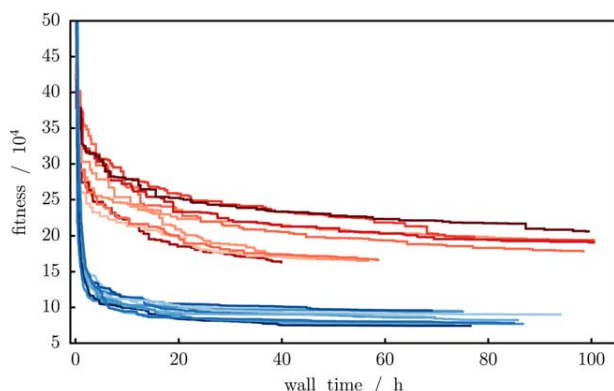


Figure 5. 10 selected but representative GA-runs with our old code (red lines)^[34,35] for the RSSR problem, in comparison to 10 with our new OGOLEM/sPuReMD-combo (blue lines, present work). The progressions of the objective function values of the ADF based runs (“old code”) and the calculations of *run4* 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

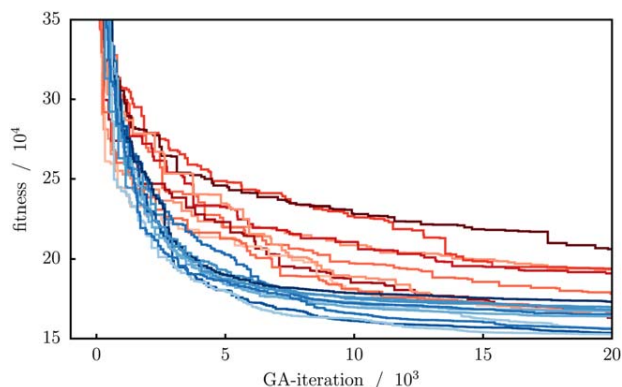


Figure 6. 10 selected but representative GA-runs with our old code (red lines)^[34,35] for the RSSR problem, in comparison to 10 with our new *OGOLEM/sPuReMD*-combo (blue lines, present work). The progressions of the objective function values of the ADF based runs (“old code”) and calculations of *run3*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

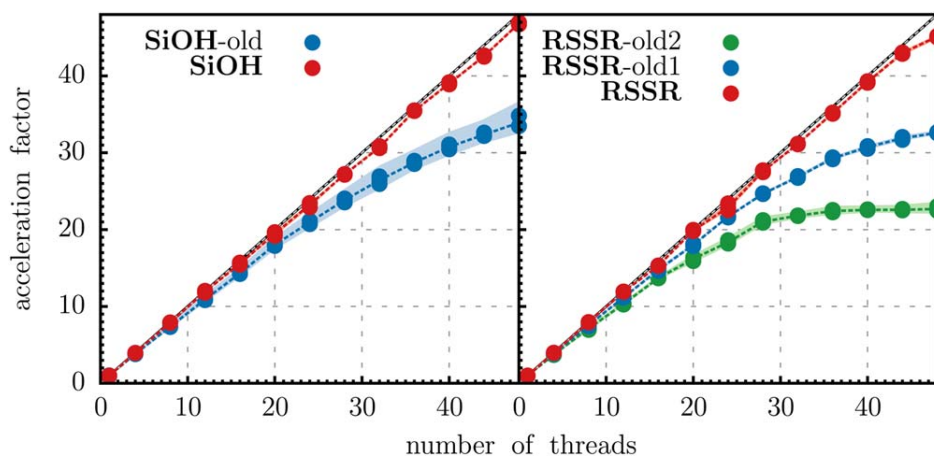


Figure 7. Strong-scaling benchmarks of the SiOH optimization problem (**SiOH**, compare Section SiOH benchmark) and a training set for the **RSSR** problem (similar to that of Section Disulfide application example) using shared-memory parallelism. Versions with “old” in their names are shown for comparison, 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. Verners, 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. Corminboeuf, *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. Schwartztruber, *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

4.2. Additional Information

4.2.1. Automation of Force Field Parameterization

The set of adjustable parameters and the reference set of the RSSR problem discussed for benchmarking purposes in our publication^[1] was designed with the idea of black-boxing the parameterization process in mind. The long-term objective of this side project was to be able to generate an optimized REAXFF parameter set just from some initial geometries with minimal effort by the user.

The following section is dedicated to the discussion of this automatization. Although it was not possible to obtain a satisfactory parameter set with the approach described here, it will be discussed in detail and the problems will be highlighted. Possible future improvements to the methodology are pointed out at the end of the section. It is worth noting that a parameter set for the RSSR system was successfully optimized in a more manual procedure. The knowledge gained during that manual optimization, which is discussed in the second part of the chapter (section 4.4.1), can be used to advance the black-boxing technique described here.

The Approach

The parameterization of a REAXFF force field with the OGOLEM software package or the older ADF-based evolutionary algorithm implementation requires three basic input sets.

The reference set contains the geometries and properties that need to be fitted during the optimization. In the current implementations the reference set is distributed to two separate files. The `geo` file contains geometries of molecules or unit cells which are identified by a unique label. The `trainset.in` file contains the reference properties for the molecular structures in the `geo` file. Examples for these input files are found in appendices A.2 and A.3 respectively.

The parameter set is that subset of empirical REAXFF parameters which are desired to be optimized. These adjustable parameters are defined in the `params` file by a unique identifier and their numerical value limits in the search space.

Finally the evolutionary algorithm need to be set up properly to yield a good approximation to the global optimum within reasonable computational time. The `*.ogo` file is used to set the calculation with OGOLEM up. It was already mentioned in the theoretical introduction that the evolutionary algorithm has a lot of input parameters itself which need to be set up properly to yield satisfying results.

Reference Set: A successful optimization revolves mostly around the quality of the reference set. Without the right amount of reference data and well balanced weighting factors for the molecular properties, the parameterization is prone to over- or underfitting.

The parameterization was aimed at medium to large organic molecules in vacuo or in solution. The dynamic properties of a molecule or reaction should be captured accurately. Therefore potential energies, gradients, equilibrium geometries and partial charges are needed as reference points.

In the spirit of the black-box approach, the reference set is desired to be built from a single input geometry or a small set of geometries along a reaction path.

The geometries for the single-point energies, which are used as reference energies for the PES, are snapshots from trajectories of the input geometry at different temperatures. The trajectories were calculated with the BOMD routine in GAUSSIAN09^[95] with the semiempirical PM6 method. To ensure sufficient coverage of the potential energy surface temperatures between 100K and 1000K were used. Snapshots were extracted from the trajectories and prepared as input for RI-MP2/cc-pVDZ calculations with the ORCA quantum chemistry package^[101,102].

The resulting energies were then entered into the `trainset.in` file with weighting factors of 1.00. The respective geometries were directly converted to the needed `biograf200` format used in the REAXFF `geo` files.

There are two main reasons for this choice of potential energy data. First, they are easily generated and handled by automatic black-box processes in arbitrary quantity. Second, the sampling of the potential energy surface is guaranteed to be dense in areas that are passed often and less dense in other, less important areas.

Assembling a representative set of geometries which features one or more specific functional groups for geometry optimizations and partial charge calculations from a single input geometry is a less trivial task. To circumvent this problem and get to results faster, the geometries were generated from SMILES¹ strings via a PERL script. The program was supplied with a starting string that contained blanks for residues. The residues were also predefined as SMILES strings and all blank connection points of the base string saturated with one residue each. The script prepares all possible residue combinations as input for a geometry optimization and partial charge analysis with ORCA^[101].

¹SMILES: Abbreviation for Simplified Molecular Input Line Entry System. Molecules may be encoded and used as ASCII strings^[103]

4. ReaxFF Parametrization and Disulfide Mechanochemistry

When all optimizations are done, the crucial information is extracted from the output files of converged calculations and transferred to the reference set. The geometries are converted into a z-matrix format by a custom PERL skript that uses OPENBABEL^[104,105] as a backend and then entered to the `trainset.in` file. The weightings of the bond lengths were set to range between 0.01 and 0.05 reflecting different degrees of relevance. The less relevant C-H bonds were assigned weights of 0.05 while more important bonds between heteroatoms got the weight 0.01. This was done to account for the much higher number of C-H bonds in bigger organic molecules compared to heterobonds and the fact that in reactions the chemist is usually more interested in C-C, C-X or X-X bonds where X is a heteroatom. Angles and dihedrals were assigned a constant weight of 3.00. All partial charges were weighted with the factor 0.01

Although in the test calculations many steps were carried out manually, the process can be abstracted into scripts almost completely. The user would only be required to set the SMILES template and residue strings, as well as forwarding the input files for the geometry optimizations to a proper HPC-machine.

A reference set generated the way described above is easy to extend or shrink if needed. Unfortunately a reference set which is convenient to build is not necessarily a good set for the optimization as the following sections show. There is furthermore no reliable way to determine the ideal size of the reference set. It is obvious that more reference data is needed the more parameters are optimized, but there is no definitive measure how many additional reference items are needed per parameter. As the quality of a parameterization generally improves when more reference data is provided, the reasonable approach to the size of the reference set should be to take as much data as can be afforded. This should minimize underfitting issues, overfitting is then taken care of by the regularization measures discussed in our publication^[3].

Parameter set: The parameter set specified in the `params` file is also critical for the optimization. The challenge is to choose the most important parameters and set reasonable boundaries in the search space.

The initial not ideal approach which was taken to identify important parameters was to use the optimized geometry of a molecule and calculate an objective value for just this geometry in the reference set. Since the underlying potential energy surface determines the molecular structure all parameters and not just the geometry parameters should be addressed.

Following this initial objective value evaluation, each empirical REAXFF parameter

4. ReaxFF Parametrization and Disulfide Mechanochemistry

was varied slightly and a new objective value was calculated. The deviation of each varied value and the base value was calculated by using equation 4.1. Parameters that influenced the objective value were categorized as sensitive parameters.

$$\Delta f_{\text{obj}} = \frac{|f_{\text{obj,base}} - f_{\text{obj,var}}|}{|f_{\text{base}}|} \quad (4.1)$$

The procedure then allowed to generate a params file which contains the most important empirical parameters. The importance was measured as the effect on the objective value. Either a cutoff can be chosen, or all parameters which influence the error value are used.

The problem with that procedure is that parameters which contribute only weakly to the equilibrium geometry may become very important along the reaction path. Using a cutoff can then exclude parameters from the optimization which are important for a reaction when only a single equilibrium geometry reference is used. This leads to underfitting.

This problem can be avoided since the evaluation used the same framework as was used for the evolutionary optimization, which made it possible to use multiple structures and relative energies for the evaluation of the most important parameters. In the end multiple structures and reference energies along the disulfide dissociation curve were used to identify sensitive parameters. Doing this is advised for any reparametrization to avoid the above mentioned problems.

The boundaries of the parameters in the search space are calculated by adding percentual margins to the current parameter value. Small margins of $\pm 5\%$ are less prone to fitting artifacts that result from exploring areas with bad parameter settings. Larger margins of $\pm 20\%$ on the other hand explore larger areas of the parameter space for an optimal solution.

The `params` file used in the benchmark calculations in the publication was initially generated with a script as described above. After that some manual adjustments were done. This was mainly because the problem set was desired to be compatible with OGOLEM and the older ADF based implementation for comparison of the two. The older implementation is less resilient to misaligned parameters und unconverged items in the reference set than the OGOLEM implementation. Therefore the `params` file had to be modified to prevent optimization runs from terminating early.

Evolutionary algorithm setup: It was already mentioned in the theoretical section 2.3.1 that the ideal operator settings are usually not known prior to the calculation.

When a black box routine for the global optimization of REAXFF parameters is to be devised, an operator setup which performs well for both exploration and exploitation is needed.

Several different operator setup have been tested for their performance. Each screened operator setup was used in ten independent calculations with a population size of 500 solution candidates and 10000 iterations each. The performance of each setup was judged on the basis of different quality measures. The measures were the best final objective value and the average final objective value out of the ten calculations as well as the rate of decay of the objective value measured as the decaying constant of an exponential fit to the best solution candidate fitness throughout the calculation.

Results

The best solution candidate found with the operator setup has an objective value of 72407. This total value decomposes into the following contributions: 65568 from the energy items, 3352 from geometries and 3487 from the partial charges section. Gradients and other properties were not used in this iteration of the reference set.

The error contributions from the potential energy section of the reference set was distributed over 487 items. The average deviation of the potential energy from the MP2 reference value is therefore 11.6 kcal/mol. The smallest deviation in the set is 0.05 kcal/mol and the largest deviation is 38.8 kcal/mol. These are very large deviations and would not be acceptable for a parameter set which is to be applied for production calculations.

It might be argued that such large deviation could be neglected if they occur on regions of the PES which are not relevant for the parameterized reaction. However, due to the random sampling by which the reference points are generated it would not be easy to decide where on the PES the problematic structures are located. Certainly it can not be decided by a black-box on the basis of the error sum alone.

It was furthermore not clear what quality of the structures made them problematic during the optimization. Over various optimization runs there was a certain consistency to the error, meaning that energy items with low errors tend to produce low errors in all runs. The same goes for the bad items. However, the error for single reference items still vary over a wide range in different optimization runs. One of the better items for example had an average error of 6.2 kcal/mol over ten optimization runs with the lowest deviation being 0.1 kcal/mol and the highest deviation being 26.7 kcal/mol. Although the differences between the runs are severe, this error range still remains clearly

4. *ReaxFF Parametrization and Disulfide Mechanochemistry*

separated from the worst reference items. An example for a high error geometry yielded an average error of 45.8 kcal/mol out of the same ten runs as the low error item before. The minimum error was 36.2 kcal/mol and a maximum error was 53.9 kcal/mol. These variations did occur independent of the final objective value reached in the optimization runs.

There are most likely two reasons for the above mentioned deviations of the reference energy and the empirical potential energy. First, the results indicate that the energies are underfitted. This may be either due to the quite restricted searchspace which was chosen for compatibility reasons or because the number of iterations was too small for the EA to converge to a satisfying solution. This hypothesis is supported by the systematic overestimation of BDEs which are discussed later. Second, some structures are for yet unknown reasons problematic for the REAXFF potential functions. The reason may also be a severe underfitting problem which could have resulted from an insufficient number of adjustable parameters in the `params` file. Another explanation may be that the MP2 wavefunction is defective in regions of the PES that are reached during the course of the PM6 trajectories.

The main contributions to the geometries error sum stem from underestimations of carbon sulfur bond lengths and connected angles. The problem was addressed in later parameterizations and is due to underfitting.

The molecular partial charges deviate by up to 0.35 from the reference ESP charges. These errors mostly occur in strongly polarized heteroatoms. The quality of atomic partial charges is always questionable, there is a multitude of calculation schemes to calculate partial charges and which partition scheme or fitting routine yields the most justifiable results is not easily decided. Therefore both, the MP2 ESP charges as well as the QEq charges yielded by REAXFF may be of poor quality.

These results mainly concern the performance of the optimization itself. The following sections will discuss the optimized parameter sets with respect to their MM results.

MD simulations with the parameter set, found by the procedures discussed above, show erroneous behaviour. Even small sets of trajectories feature unexpected hydrogen transfer reactions and subsequent dissociation which should not occur. Again this shows that the empirical parameters are not fit to be used in production simulations.

A more positive view on the matter is that the MD simulations yielded stable trajectories at different temperatures and simulation times of 25-100 ps. This is not guaranteed for global optimization runs on REAXFF parameters. Trajectories from other optimization runs often featured spontaneous atomizations or literally crumpled molecular

geometries which were also reported before by *Larsson*^[100].

The potential energy curves in figures 4.1 and 4.2 reveal further qualities of the reparameterized potential.

Compared to the parameterization by *Mattsson* and coworkers, which was used as a starting point for the optimization, the potential curves for diphenyl disulfide (DPDS) compare well to the CASPT2 reference curves at the potential minimum. This good agreement occurs even though no explicit DPDS energy data is found in the reference set.

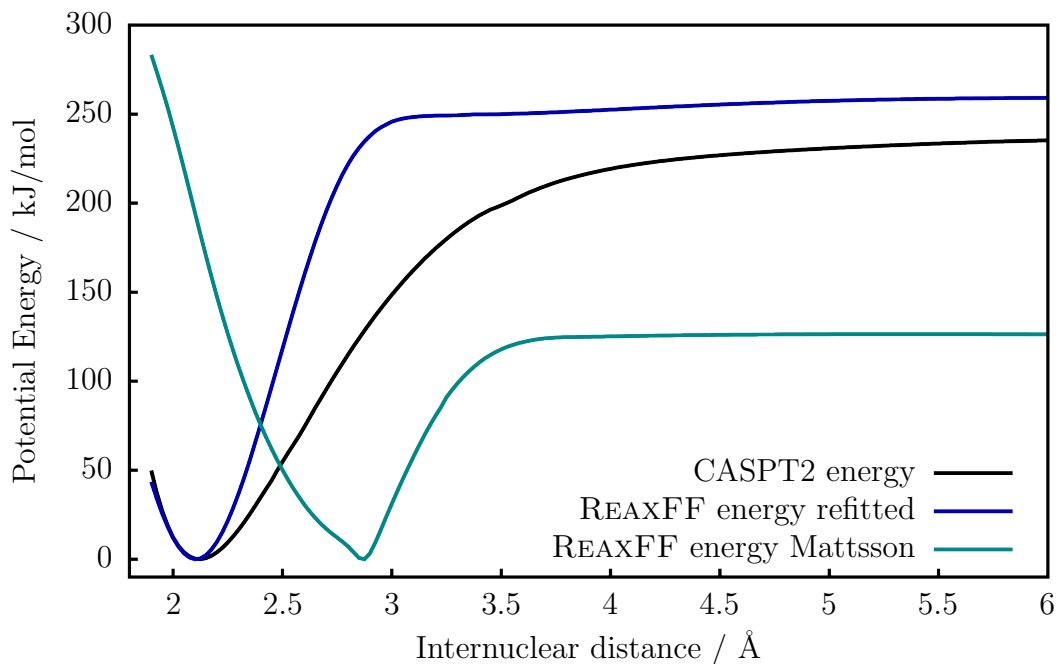


Figure 4.1.: Potential energy curves for the SS bond distance coordinate in DPDS.

For higher internuclear separations the REAXFF potential curve starts to differ drastically from the CASPT2 reference. This is even worse for the dissociation potential of the carbon sulfur bond in DPDS. The potential curve in figure 4.2 features a BDE that is off by a factor of two. Furthermore the curve has an artificial maximum along the dissociation and an unphysically high gradient.

Other degrees of freedom show similar behaviour, the accuracy near the equilibrium is high but becomes increasingly deficient when altering the geometry of the system. The occurrence of this phenomenon is assumed to be related to the choice of reference data. As said above, the sampling of the potential energy surface takes place mainly around the potential minimum while high energy regions are not probed at all. This leads to

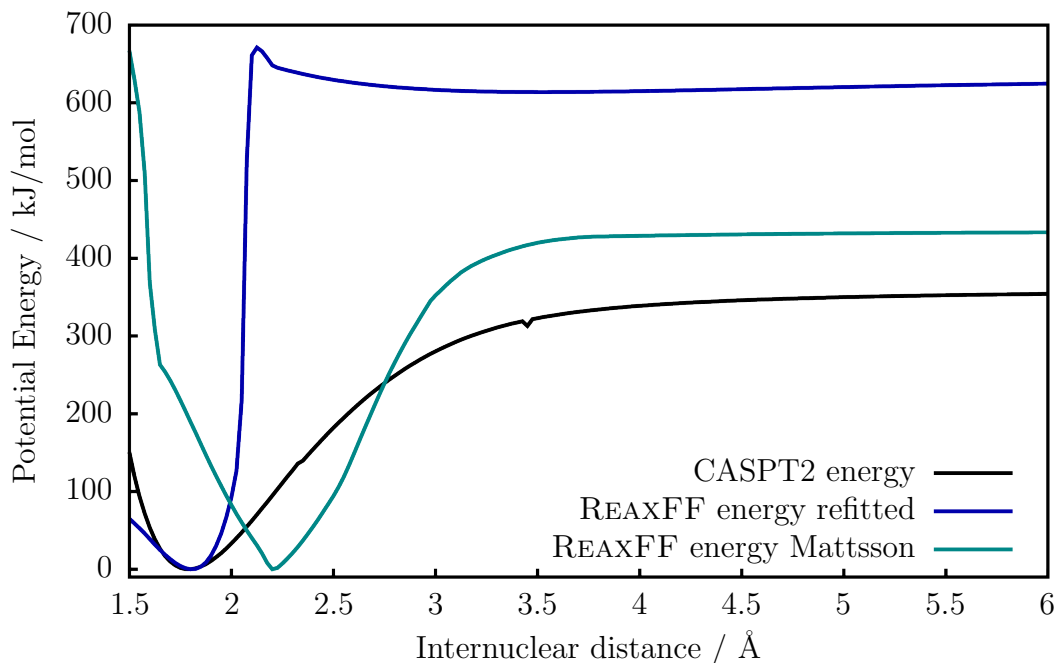


Figure 4.2.: Potential energy curves for the CS bond distance coordinate in DPDS.

an imbalance of the reference data which is skewed towards low energy configurations. The subsequent optimization will fail to yield parameters which can predict high-energy structures like transition states or dissociation products correctly since the reference set contains no information about them.

The error results for the energy section of the reference set during the EA may also be partially explained by this behaviour. It is suspected that the structures with high energy errors have some bonds lengthened to a degree where they lie in the very erroneous region of the REAXFF PES.

Further Work

The approach outlined above has various flaws and, as the results indicate, much room for improvement.

The optimization itself can be improved in various ways. The operator settings were tuned for maximal exploitation of objective function funnels and the number of iteration was too low to yield converged results. Regularization procedures as used in machine learning should be used to avoid overfitting issues. An early stop procedure can be easily implemented for a generic reference set by taking a randomized subset of the

reference set as comparison set. For future production runs, the operator settings and iteration number should be set according to the way discussed in length in our later publication on the matter^[3] which is discussed in the next section.

In the 2016 publication further problems concerning the `params` file were addressed. Obviously some of the above mentioned problems are due to underfitting and ill-chosen parameter boundaries. Increased knowledge about the parameters can be applied to generate improved `params` files. Most of the details regarding the parameters and search space setup are discussed in length in the next section.

The most vital part of the optimization still is the reference set. The first incarnation of the reference set, which was described above, lacked information about the high-energy regions of the PES, especially dissociation asymptotes. Therefore other methods to sample the potential energy surface need to be employed which also visit the regions that are left out by normal BOMD trajectories. Possible approaches would be to use steered molecular dynamics to sample all of the accessible PES more efficiently or following normal coordinates until the potential energy becomes constant. Both approaches could be implemented in a black box fashion, however, wave function methods will be hard to converge in the regions of the PES where one or more covalent bonds are broken. This means the automatic exploration of the PES for reference points will lose some of its black-box character.

This iteration of the reference set furthermore lacks gradients as properties which were newly introduced in the OGOLEM implementation. It is technically unproblematic to obtain also gradients for a predefined portion of the trajectory structures in the reference set. Fitting the gradients may soften the drastic deviations of the potential energies.

4.2.2. Gradients as Additional Property

With the migration of the evolutionary optimization from the ADF-based implementation to OGOLEM done by *Dittner* and *Aktulga*, molecular gradients were introduced as new reference property which may be fitted during the parameterization. This was deemed necessary since the global optimization had problems to smooth out the potential energy curves of bonds with allowed orders of two and three.

A misalignment of bonding parameters leads to situations where additional artificial minima occur along the potential energy curve. Those curves may fit the reference energies rather well, but the wrong sign of the gradient will lead to explosions in MD calculations.

An example for this behaviour is depicted in figure 4.3. The potential energy curve

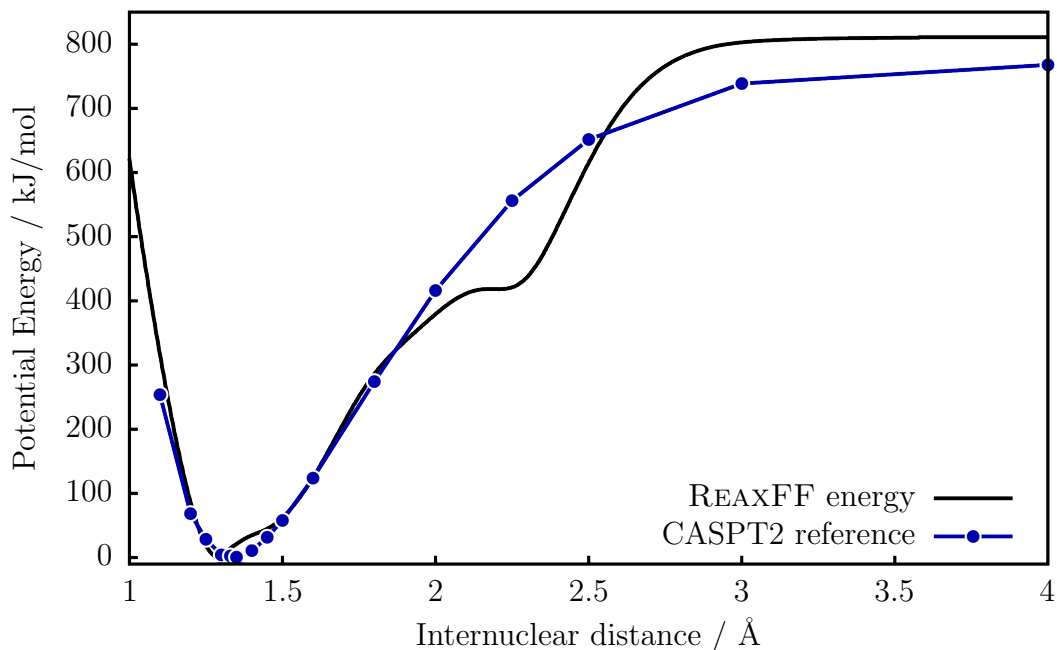


Figure 4.3.: Potential energies for the carbon-carbon double bond in ethene. The CASPT2 reference points (blue) are connected to guide the eye.

shows the stretching along the carbon double bond coordinate in an ethene molecule. While the energy deviation and therefore the error contribution is relatively small, the potential curve is qualitatively wrong compared to the reference curve.

At the point of the artificial minimum at 2.2 Å the gradient of the potential energy curve should be large and positive and the strict increase of the energy from the equilibrium distance to the dissociation asymptote does not allow for negative gradients in this region at all. Therefore the qualitative defect can be penalized by the objective function by supplying the molecular gradient at this point.

4.3. Publication: ReaxFF Reactive Force Field for Disulfide Mechanochemistry, Fitted to Multireference ab Initio Data.^[3]

Contribution to the paper:

- Main contribution to the publication text.
- Assembly of reference sets consisting of MP2 and CASPT2 data.
- Curation of `params` files for the optimization.
- Optimization and assessment of the REAXFF parameter set.
- Proof-of-principle MD calculations for strained mechanophores in vacuo and in solution.

Full text reprinted with permission. Copyright 2016 ACS Publications.

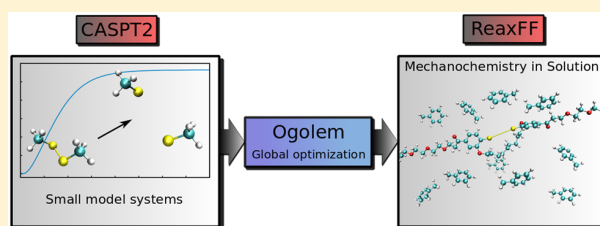
REAXFF Reactive Force Field for Disulfide Mechanochemistry, Fitted to Multireference *ab Initio* Data

Julian Müller and Bernd Hartke*

Institute for Physical Chemistry, University of Kiel, Olshausenstrasse 40, 24098 Kiel, Germany

S Supporting Information

ABSTRACT: Mechanochemistry, in particular in the form of single-molecule atomic force microscopy experiments, is difficult to model theoretically, for two reasons: Covalent bond breaking is not captured accurately by single-determinant, single-reference quantum chemistry methods, and experimental times of milliseconds or longer are hard to simulate with any approach. Reactive force fields have the potential to alleviate both problems, as demonstrated in this work: Using non-deterministic global parameter optimization by evolutionary algorithms, we have fitted a REAXFF force field to high-level multireference *ab initio* data for disulfides. The resulting force field can be used to reliably model large, multifunctional mechanochemistry units with disulfide bonds as designed breaking points. Explorative calculations show that a significant part of the time scale gap between AFM experiments and dynamical simulations can be bridged with this approach.



1. INTRODUCTION

With the introduction of the atomic force microscope (AFM) in 1986 by Binnig and co-workers,¹ atom-scale probing of surfaces and molecules became possible. This analytical tool proved to be very handy for the investigation of mechanochemical processes. The field of mechanochemistry emerged in the beginning of the 20th century and contains fields such as tribology, sonochemistry, and single-molecular experiments. The latter is now accessible for some ten years due to the development of sophisticated AFM experiments.^{2–4} Hence, experiments can now probe force-dependent compound lifetimes, conductivities, and other observables.

In contrast, theoretical approaches to mechanochemistry have to deal with several problems: While static properties can be described well with approaches like COGEF⁵ or EFEI⁶ within quantum chemistry,⁷ dynamic simulations of mechanochemical events face severe difficulties: Unless the typical polymer ends and anchor groups needed to put the molecule between AFM tip and surface are cut off completely, the system size is too large for standard first-principles molecular dynamics. The density functional theory (DFT) level typically used in such calculations may be unable to describe the desired bond dissociation sufficiently accurately; and even if all these problems could be surmounted, at such a level of theory there definitely is no chance of bridging the >12 orders of magnitude between the subfemtosecond time steps needed and the millisecond-to-second time scale of AFM experiments. Force-field-based molecular dynamics (MD) has now arrived at routine millisecond simulations,⁸ but off-the-shelf force fields often do not allow bond dissociations by construction.

Reactive force field (FF) methods can be a way out of this dilemma. By employing direct relations between the molecular

geometry and dynamic bond orders of atoms, reactive potentials have become more stable, more flexible, and therefore more widely used. Two bond-order potentials that have been applied to a whole array of different problems in recent years are COMB and REAXFF.^{9,10} Other reactive FF formalisms that implement novel concepts, for example DBO-FF¹¹ or QCT-FF,¹² still have to be tested more broadly. In this work, we focus on the use of REAXFF. The REAXFF formalism that was introduced by van Duin in 2001¹³ allows for the calculation of big systems and boxes with periodic boundary conditions with a speedup of about 10000, compared to DFT calculations. With this speedup, it is possible to reach time scales and system sizes^{14,15} that are needed to investigate, e.g., AFM-induced unfoldings of macromolecules in solution.

So far, no reactive FF, including REAXFF, can combine true universality with sufficient accuracy. The latter can be achieved, however, by fitting the reactive FF to a restricted choice of systems and/or reactions, in our case to mechanochemical conversions of disulfides. This fitting has to be repeated for different system and reaction choices. Each fitting task is difficult: There is an estimated number of 80 ± 20 relevant parameters per atom that need to be fitted against reliable reference data to obtain a force field capable of capturing the problem at hand. Therefore, efficient parameter optimization techniques are vital for reactive FF use. Various optimization approaches were applied by different groups. For REAXFF, most of the work groups use van Duin's SOPPE routine¹⁶ to fit parameters for various applications.^{17–19} Multiobjective evolutionary strategies^{20,21} as well as simulated annealing²² have also

Received: May 5, 2016

Published: July 14, 2016

been used for this task. In this work, we apply the evolutionary algorithm (EA) strategy established in previous publications of our group.^{23–25}

The motivation for the parametrization of a force field for disulfide mechanochemistry was provided by the single-molecular experiments by Schütze et al.⁴ According to DFT calculations in our group, the triazole macrocycle used there presumably has nearly identical activation forces for the retro click reaction and for other bond fissions. In contrast, a disulfide bond should be much easier to break. Dissociation events therefore should be much simpler to distinguish in disulfide-containing mechanophores, and rupture events may be better resolved during AFM extension experiments. The resulting sulfide biradicals might even recombine without the presence of a catalyst, if they can be brought into close proximity to each other again after the fission. A further possibility includes connecting the central disulfide moiety to conductive polymers. For the reasons mentioned, the two mechanophore systems shown in Figure 1 were targeted by our

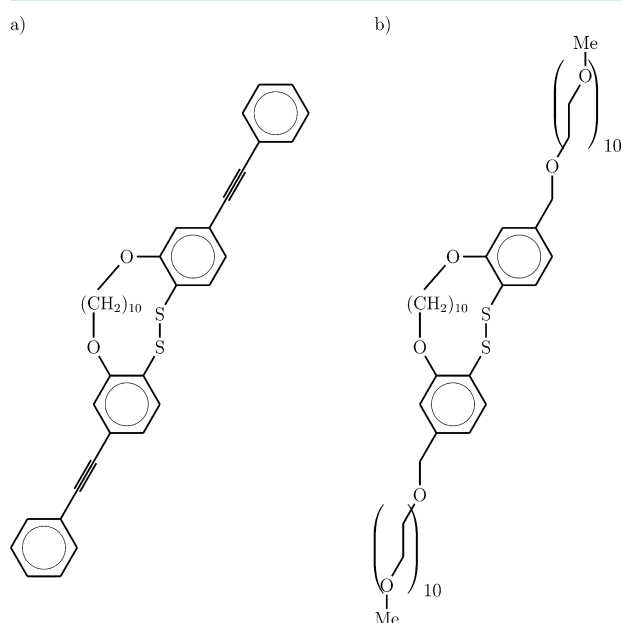


Figure 1. a) (DSM-C) Disulfide containing macrocyclic mechanophore planned to be subjected to conducting AFM experiments in toluol. b) (DSM-PEG) Disulfide containing macrocyclic mechanophore for retraction experiments in vacuo or in air.

collaboration partners (work group Lünig, Christian-Albrechts-University Kiel) via organic synthesis and hence were also investigated with the newly parametrized force field in the present work. The disulfide mechanophore in panel (a) is intended for conducting AFM experiments in solution and will be dubbed DSM-C. The second structure (panel b) is functionalized with polyethylene glycol chains and will be called DSM-PEG.

Based on the EA parameter fitting strategy mentioned above, we present a reparametrized REAXFF force field for the mechanochemical description of the disulfide moiety shown in Figure 1. To our knowledge, so far only one single publication²⁶ has pointed out this link between REAXFF and mechanochemistry, again for the case of disulfide bridges in proteins. There, however, the objects of study were the

combined influences of various redox agents and of mechanical strain on the stability and reactions of disulfide bridges in proteins. For this purpose, a previously published parametrization of REAXFF for proteins (specifically, for glycine²⁷) was used, which employed DFT data as reference. In contrast, in the present work, we target AFM mechanochemistry of designed nonprotein organochemical compounds, in the absence of additional chemical influences that affect the disulfide bond. This requires a dedicated refitting of REAXFF, as we show below. In addition, we depart from the defacto standard of using only DFT data as reference. It is well known that breaking (and making) of covalent bonds, and even of only strongly stretched covalent bonds, cannot be described with quantitative accuracy by effective single-particle/single-determinantal models like HF and DFT. For the present application example, this is demonstrated in detail in section 3.1. Hence, to correctly capture the covalent chemistry involved in the mechanical activation of disulfide bonds, we employ ab initio multireference perturbation theory results (CASPT2) as sulfur–sulfur reference data into our REAXFF parametrization.

2. THEORY

2.1. REAXFF. The REAXFF formalism that was introduced by van Duin et al. in 2001¹³ is a bond order potential, i.e., all bonded interactions are based on bond orders that solely depend on atomic coordinates. In the most recent version of the formalism,²⁸ the bond order between two atoms takes the form

$$\begin{aligned} BO'_{ij} &= BO'_{ij}{}^{\sigma} + BO'_{ij}{}^{\pi} + BO'_{ij}{}^{\pi\pi} \\ &= \exp\left[p_{\text{bo}1}\left(\frac{r_{ij}}{r_0^{\sigma}}\right)^{p_{\text{bo}2}}\right] + \exp\left[p_{\text{bo}3}\left(\frac{r_{ij}}{r_0^{\pi}}\right)^{p_{\text{bo}4}}\right] \\ &\quad + \exp\left[p_{\text{bo}5}\left(\frac{r_{ij}}{r_0^{\pi\pi}}\right)^{p_{\text{bo}6}}\right] \end{aligned} \quad (1)$$

Here the prime indicates that these are uncorrected bond orders. They are later modified to account for over- and undercoordination that occurs by geometrical distortion or chemically unusual coordination. The bonded energy contributions to the system are then calculated using the corrected bond orders. For example the covalent bonding energy equates to

$$E_{\text{bond}} = -D_e^{\sigma} BO_{ij}^{\sigma} \exp[p_{\text{be}1}(1 - (BO_{ij}^{\sigma})^{p_{\text{be}2}})] - D_e^{\pi} BO_{ij}^{\pi} - D_e^{\pi\pi} BO_{ij}^{\pi\pi} \quad (2)$$

The total energy of the system is a sum of contributions from covalent bonds, lone pairs, over- and undercoordination, angles, penalties and conjugation terms, torsions, hydrogen bonds, van der Waals interactions, and Coulomb terms.

$$\begin{aligned} E_{\text{tot}} &= E_{\text{bond}} + E_{\text{lp}} + E_{\text{over}} + E_{\text{under}} + E_{\text{angle}} + E_{\text{pen}} + E_{\text{coa}} \\ &\quad + E_{\text{C}2} + E_{\text{tors}} + E_{\text{conj}} + E_{\text{H-Bond}} + E_{\text{vdW}} + E_{\text{Coulomb}} \end{aligned} \quad (3)$$

For further details on REAXFF, we refer to the original papers^{13,17–19,27} and reviews.^{9,10}

From the functional form of REAXFF as seen in the equations above it is evident that the flexibility of the potential energy surface comes at the cost of a manifold of parameters that need to be fitted to reliable reference data. Depending on the extent of reparametrization planned, somewhere between 50 and 300

parameters form the space that needs to be explored for the best fitting parameter set. Since the structure of this search space is so extremely challenging²⁵ that local optimization methods run a high risk of getting stuck in bad minima, global optimization in the form of evolutionary algorithms was employed.

Global Optimization. Force field parameter values have to be adjusted such that force field results reproduce all desired properties of a system. Since it is unrealistic and undesired to calculate all that data prior to the optimization, they are fitted to a subset of all possible data called a reference set. Such a reference set may consist of quantum chemical data, experimental results, or a combination of both. However, a measure is needed to assess the quality of a parameter vector. As a measure of quality a sum of squared deviations is used. The global minimum of this objective function (eq 4) is probably the ideal set of parameters, but low-deviation locally minimal solutions are likely to be acceptable in practice, too. However, low deviation is not the only criterion that needs to be tracked. Over- and underfitting are serious problems that have to be avoided, as explained below.

$$F_{\text{obj}}(x_i) = \sum_i \frac{(x_i - x_{i,\text{ref}})^2}{w_i^2} \quad (4)$$

The variables x_i in eq 4 are the molecular properties used to parametrize the force field. In this work, relative energies, molecular structures, and gradients were used as reference for the parametrization of the REAXFF energy surface. The relative importance of each reference set item is adjusted by weighting factors w_i .

The number of local minima of this objective function grows exponentially with its dimensionality. However, from previous experience on this^{23,24} and similar problems,²⁹ we expect that there will be many low-lying minima scattered across search space but both with values of the objective function and with performance outside of the reference set similar to each other and to the global minimum. Therefore, evolutionary algorithms (EA) are a method of choice, since they combine a quick initial approach to promising regions with a population of solutions to choose from. An obvious choice of an EA package is the software suite OGOLEM that was developed in our work group,^{30,31} which offers efficient parallelization by construction, due to a pool-based EA-structure.³² In previous work,²⁵ Dittner already interfaced this OGOLEM suite with a highly efficient C reimplementations of REAXFF by Aktulga,³³ and it was demonstrated that the resulting software combination was then able to globally optimize REAXFF parameter sets with high parallel efficiency.

The applied evolutionary algorithm is pool-based.³² As described in this reference, the original aim of the pool algorithm was better parallel efficiency. In the present context, however, a second advantage arises: If a new individual with a better fitness than that of the weakest individual in the pool is found during the evaluation, the new individual is added to the pool and the worst individual is discarded. No individual can be accepted in the pool that has a worse fitness than the weakest individual in the pool, a behavior that can arguably be called the pool version of elitism. Since fitness evaluation according to eq 4 is a sum accumulated in many small steps that can only contribute positive values, this elitism can already be anticipated during fitness evaluation. The so-called “immediate fallback” feature was introduced by Dittner²⁵ and consists in

stopping the objective function evaluation as soon as the fitness reaches the threshold of the worst individual. Since every contribution to the sum in eq 4 costs computer time, immediate fallback incurs substantial time savings. As a convenient additional feature, these savings are largest for the worst individuals; in other words, computational expense is focused automatically on the best individuals. The fitness evaluation itself is a blackbox routine to the EA, which receives a candidate vector of parameters and returns an objective function value.

The ability of an EA to quickly find promising regions of search space has the downside that individuals from the best region found so far may take over the whole EA pool. This is termed premature convergence, and it clearly weakens the exploratory power of the EA toward possibly even better regions. This effect can be prevented by niching, which has a long history in EAs (cf. ref 34 and references cited therein). With niching, only a definite number of individuals from a certain confined region of the search space, called a niche, is allowed in the pool. For the realization of niching, the search space is divided into hypercuboid subspaces of adjustable size. In this coarse grained space, nearness of two individuals is defined by the number of subintervals shared by them. To set up the niching, two further parameters are relevant: One is the number of individuals that are allowed in one niche, and the other is a measure for how “near” two individuals need to be to belong to the same niche.

In the present work, it turned out to be reasonable to allow exactly two individuals in one niche. The very rugged search space (illustrated in ref 25) and the optimization power of the EA operators employed lead to the phenomenon that individuals in the same niche tend to collapse to almost identical solutions rather quickly. Therefore, allowing more than two individuals in a niche would be a waste of computing resources. On the other hand, allowing for only one individual would result in less efficient exploitation of local minima. Hence, two individuals per niche are a good compromise.

The allowed nearness criterion should be chosen such that a reasonable balance is achieved between maximizing exploration and maximizing computational savings via the immediate fallback feature. The narrower the niches and the bigger the pool, the worse will the fitness of the weakest individual be. Bad individuals are not a direct aim, but they do enhance exploration. However, immediate fallback incurs greater computational savings the better the fitness of the weakest individual is.

In every step there are two vectors chosen to be subjected to either mutation or crossover. The chance to choose either one can be set by the user. The choice of the parent vectors is based on probability distributions with adjustable widths. Obviously, the balance between exploitation and exploration can be influenced by adjusting these widths. From the many variants of crossovers implemented in OGOLEM, two were found to be the best for global optimization of force fields. One of these is a multipoint exchange operator with cutting points randomly distributed along the vector. The second operator is a mixing recombination operator that generates child alleles as a weighted average of the parents values.

From the possible mutation operators, again two realizations work best. The first variation generates new values for a number of alleles that are linearly distributed over the search space. The second option is to generate Gaussian-distributed

values around the current allele value or around the center between the parameter boundaries.

The different typical setups used in different stages of the present work are shown in Table 1.

Table 1. Setups of the Evolutionary Algorithm As Used for This Work^a

EA setups used		
exploration setup	poolsize	2000
	steps	2.500.000
	parents choice	wide
	80% crossover	80% multipoint exchange 20% mixing operator
	20% mutation	40% Gaussian, wide, center 40% Gaussian, medium, current 20% linear operator
exploitation setup	poolsize	500
	steps	early stop
	parents choice	narrow
	70% crossover	80% multipoint exchange 20% mixing operator
	30% mutation	40% Gaussian, medium, current 40% Gaussian, narrow, current 20% linear operator
local relaxation	poolsize	seeded pool
	steps	as many as affordable
	parents choice	all
	100% mutation	100% Gaussian, very narrow, current

^aThe exploration setup was applied during the assembly of the reference set, to be able to sample vast regions of the search space. The exploitation and local relaxation setups were chosen for the final optimization or for other test runs where strongly optimized vectors were needed.

3. FORCE FIELD FITTING

3.1. The Need for Multireference *ab Initio* Reference

Data. As pointed out at the end of the Introduction, single-reference quantum chemistry methods (including present-day standard DFT) are unable to treat homolytic bond breaking accurately. This general textbook claim can be readily illustrated for the present application case.

Figure 2 depicts the energy of dimethyl disulfide as a function of the S–S distance, for three different approaches: a GGA functional, a hybrid functional, and CASPT2. Obviously, all three methods agree fairly well on the equilibrium distance, but nevertheless both DFT curves arrive at a dissociation asymptote that is in error by 45 kJ/mol (or 16% of the total dissociation energy). The CASPT2 asymptote matches the experimental value of 283 kJ/mol³⁵ perfectly, while the DFT results are far off. This gets even worse for diphenyl disulfide, with deviations up to 83 kJ/mol (or 55% of the total dissociation energy), but then the CASPT2 calculation becomes hard to converge with respect to basis set and active space. Additionally, in-between the equilibrium distance and the asymptote, slope and shape of these curves differ and depend on the choice of the DFT functional.

The reason for these differences is the qualitative change in the electronic wave function. In the vicinity of the equilibrium distance, a single-determinant approximation works very well; this is the prescribed wave function form in standard DFT. However, this approximation breaks down severely at longer

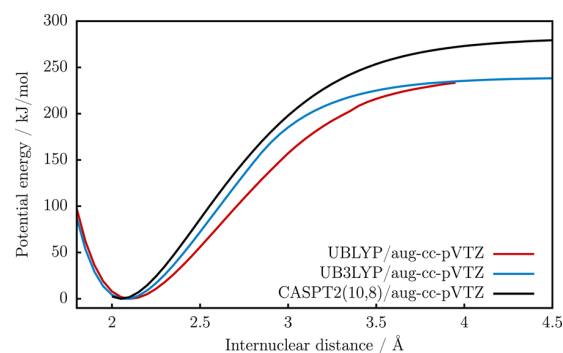


Figure 2. Potential energy of DMDS plotted over the internuclear distance between the sulfur atoms. The CASPT2 reference shows qualitatively different dissociation energies than the DFT potential curves. Additionally the curvature of the potential along the dissociation coordinate varies for different functionals.

distances, not only in the dissociation asymptote but already before. This is qualitatively illustrated in Figure 3 by the weighting coefficients of a few important contributions to the total wave function in the CASPT2 calculation. While the coefficient of the HF determinant is close to 1.0 near equilibrium, this value drops down to 0.7 in the asymptote. In the distance region where the maximum force is to be expected (highlighted in the figure), the HF determinant coefficient still is between 0.8 and 0.95, but this is not really an indicator for validity of the single-determinant picture: As the other two weighting coefficients show, the character of the wave function changes drastically just in this region. This is impossible to represent in a single-determinant approximation.

In mechanochemistry, forces from bond stretches are probed directly, i.e., first derivatives of the curves shown in Figure 2. Of course, this magnifies the differences between those curves and reveals further problems, cf. Figure 4. Not only are there differences of up to 23% between the maximum forces, but the changes in wave function character are echoed as kinks and

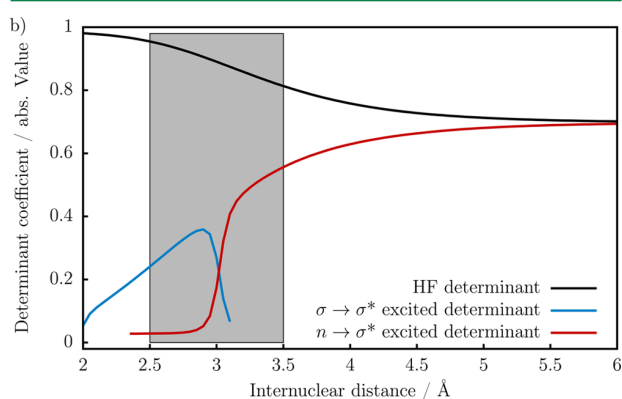


Figure 3. Weights of three important determinants for the multireference wave function of DMDS at different internuclear separations of the sulfur atoms. The mechanochemically important region of the steepest increase of the potential is highlighted in gray. While the Hartree–Fock ground state is still dominating the wave function in this region, other determinants with significant weightings are contributing to it, and in strongly changing amounts. This necessitates the use of multireference methods to yield reliable results for this region of the PES.

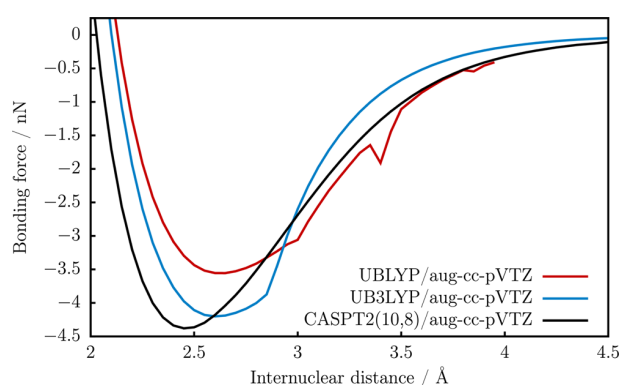


Figure 4. Negative gradients of the DMDS potential energy curve for the sulfur dissociation coordinates in DMDS. The CASPT2 reference shows a higher critical force than seen in the DFT estimates. Furthermore, the density functional curves are showing qualitative artifacts in the position of the critical point and the course of the curve at the region of the avoided crossing between the ionic and the homolytic dissociated solutions.

convergence problems in the DFT calculations, while the CASPT2 forces are smooth throughout.

With an in-depth search in the DFT functional zoo, it may be possible to come up with a functional that matches the CASPT2 data for DMDS S–S dissociation more closely than the functional choices made in this subsection. However, this will then be due to fortuitous error cancellation. Such a cancellation cannot be trusted to hold across all the necessary variations in geometries and molecular systems. Hence, CASPT2 simply is better suited to this problem and offers the needed robustness. This approach can be chosen as *REAXFF* reference level, since nothing in the *REAXFF* setup is directly tied to DFT data as reference.

3.2. Reference Set. In the overwhelming majority of published *REAXFF* papers so far, DFT data were used as reference. In contrast, in our case, the reference set used consists of MP2 and CASPT2 data, for the reasons discussed above. While the MP2 method was sufficient to perform the necessary geometry optimizations, multireference calculations were needed to calculate potential energies and gradients. All geometries were optimized with the RIMP2 method as implemented in ORCA,^{36–38} using an aug-cc-pVTZ basis set. A complete list of molecules used may be found in the [Supporting Information](#). Multireference calculations of potential energy curves and gradients for the systems dimethyl disulfide (DMDS), dimethyl thioether (DMTE), methyl dithiol (MDT), dihydrogen disulfide (HSSH), and hydrogen sulfide (HSH) were all performed using the MOLCAS program package version 8.0.^{39–41} All active spaces included the σ - and σ^* -orbitals of any bond broken during the rigid scans. Furthermore, nonbonding orbitals with π -character for sulfur were included in most cases, since they have a major influence on the electronic structure along the dissociation coordinates. All coordinates that were scanned for the reference set are compiled in [Table 2](#), together with their respective active spaces.

The reference set contains 12 to 23 single-point data pairs along the potential energy curves of every coordinate in [Table 2](#). Every pair consists of an energy relative to the globally relaxed structure of the respective system and a numerical gradient vector. This yields a total of 221 single points in the

Table 2. Overview of the Internal Coordinates That Were Scanned To Build the Reference Set and of the Corresponding Active Spaces (m, n) (m Electrons in n Orbitals) for the Multireference Calculations^a

system	coordinates	active space
DMDS	r(CS), r(SS), a(SSC), d(CSSC)	$\sigma_{SS}, \sigma_{SS}^*, \sigma_{CS}(x2), \sigma_{CS}^*(x2), n_s(x2),$ (10,8)
DMTE	r(CS), a(CSC)	$\sigma_{CS}(x2), \sigma_{CS}^*(x2), n_s, (6,5)$
MDT	r(CS), a(SCS)	$\sigma_{CS}(x2), \sigma_{CS}^*(x2), n_s(x2), (8,6)$
HSSH	r(SH), r(SS), a(SSH), d(HSSH)	$\sigma_{SS}, \sigma_{SS}^*, \sigma_{HS}(x2), \sigma_{HS}^*(x2), n_s(x2),$ (10,8)
HSH	r(SH), a(HSH)	full valence, (8,6)

^aThe letters r, a, and d found in the coordinate column are shorthand notation for internuclear distances, angles, and dihedrals. The orbitals in the active space are denoted by the character of the molecular orbital and of the corresponding bond. If multiple orbitals of one type are found in the space, this is indicated in parentheses.

reference set. Such a small set might appear to be prone to overfitting issues if 80 or more parameters are to be optimized, but every single point contains a gradient vector the entries of which are handled separately by the optimization algorithm. While these are not completely independent variables, this increases the true number of reference properties to 4622 nonetheless. The optimized geometries contribute another 255 entries to the reference set. Hence, in total, the reference set contains 4877 items. In addition to this de facto sizable reference set, countermeasures were applied to overcome the overfitting problem that are described below.

The weighting factors w_i for all reference energies were initially chosen to be 1.00. All remaining weights were adjusted accordingly to ensure that the contributions to the total value of the objective function are evenly distributed over the three sections geometries, gradients, and energies. In the final version of the reference set, the weighting coefficient of the HSSH dihedral energies was reduced to 0.3 since the errors in the resulting energy curve were too large to be neglected.

The unit for the gradient as used in OGOLEM is E_h/a_0 , while energies are given in kcal/mol. Hence, typical deviations in these two items have rather different numerical values, which implies different but hidden relative weightings of these two items in the objective function. To compensate for this effect, the coefficient for all gradients was chosen to be 0.01.

In the geometry section of the reference set, all geometries were provided as full z-matrices. All angles and dihedrals were assigned a weight of 3.00, while the bond lengths had coefficients between 0.05 for the less important carbon hydrogen bonds and 0.01 for the highly important bonds between two hetero atoms. Again, most of these discrepancies reflect the differences in units, in this case degree and Ångström, and the associated tolerances.

The overall effect of these weight settings, with units in mind, is to give otherwise fairly meaningless objective function values a direct interpretation: Dividing the total objective function value by the number of reference data set entries allows for a first, rough assessment of whether “chemical accuracy” (1 kcal/mol) has been reached or not, within the reference set. Of course, the actual performance of a parameter set has then to be tested further, but having a quick first quality indicator has turned out to be useful in online monitoring of optimization progress.

3.3. Validation Data. During the reparametrization process it became apparent that the reference set alone will not suffice

to produce satisfying results. Additional reference data was needed for the validation of solution vectors. To avoid confusion, the term “the validation set” will be used when referring to the specific portion of validation data used in the final early stopping optimization described below. Otherwise, “validation data” denotes any bit of calculated information outside the reference set. It was used to identify simulation artifacts produced by the solution vectors, which may have been caused by over- or underfitting.

In the first, rather extensive, phase of the reparametrization that comprised assembly of the reference set, constructing the properties of the parameter vectors subjected to optimization and setting up the evolutionary algorithm for optimal results, all intermediate solution vectors were tested against validation data manually. These tests consisted of calculations that could be easily checked by visual inspection. The reference set consisted of small model systems; therefore, it was of major interest how well the results translated to the full systems DSM-C and DSM-PEG. The results from these manual validations were important for setting up the parameters and the reference set for the optimization, because the defects seen in validation allowed us to conclude which parameters needed adjustments and which data should be added to the reference set.

The validation set used in the early stopping approach contained several geometries that were not included in the reference set and also energy data for diphenyl disulfide (DPDS). Due to the smaller basis set and a small active space, the multireference data for DPDS is of significantly lower quality than the calculations used in the reference set. Nevertheless, these DPDS CASPT2 data still are qualitatively correct, even toward dissociation of SS and CS bonds, and hence are a valuable contribution to the validation set. In contrast, in these situations, single-determinantal DFT calculations run into convergence difficulties or produce bad results. The complete validation set is also given in the [Supporting Information](#).

3.4. Overfitting. As stated above, overfitting is a problem that may easily occur when increasing the number of free parameters. In this work, various methods were used to detect overfitting artifacts and eliminate them. During the building of the reference set, all checking was done manually because in this stage of the reparametrization information on the detailed misbehavior of the surface was much more crucial than the fact that overfitting was present. When the building of the reference set was complete, the manual error detection was substituted by a semiautomated method that took the fitness of the dedicated validation set described above into account.

An indicator for the presence of overfitting is a final fitness value that is “too good”: As argued above, the weighting factors of the reference set entries were chosen in such a way that the error per entry roughly takes the value 1.00 when chemical accuracy is reached. To our experience, chemical accuracy or slightly less is a typical optimum of what can be expected using the REAXFF formalism. Hence, with our weightings, if the algorithm returns a fitness value that is significantly lower than the total number of reference set entries, this is a first clue toward possible overfitting.

Since overfitting is a serious concern in applications, let us illustrate this with an example: The optimized force field presented in this work has a final fitness value of 12393 for the reference set with 4877 entries. This corresponds to a squared deviation of about 2.54 per item, which was found to be a reasonable value over the course of many optimization runs. A

reference set used to illustrate the overfitting problem had 485 entries which all solely concern the SS coordinate in DMDS, while the same parameter vector as before was used. The evolutionary algorithm brought the objective value down to 167, corresponding to 0.35 per entry. The potential curves shown in [Figure 5](#) clearly illustrate the undesired consequences

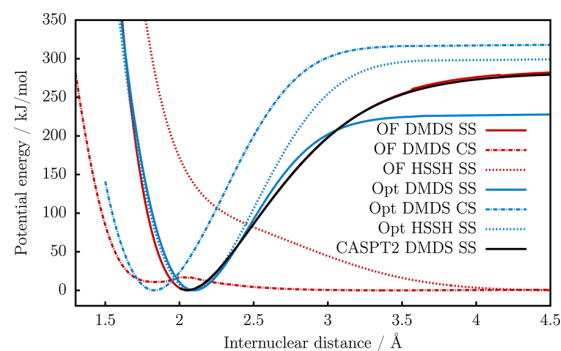


Figure 5. Potential curves for different cases: Red: overfitting (OF) issues are present, Blue: overfitting was avoided; Black: multireference data in the reference set.

of such an overfitting. While the data included in the reference set is reproduced perfectly, any other potential curve generated with this parameter set is qualitatively wrong. In contrast, in the case of the well-balanced reference set, the overall deviation of individual potential curves from the reference set is bigger than in the overfitting case, but the overall qualitative shape of the potential energy surface is captured correctly.

Another approach to assess overfitting was to test the optimized solution vectors against validation data. For example, if a well optimized force field is employed, geometry optimizations of the full system will converge to structures that compare well to the MP2 results, and molecular dynamics calculations of the system will show the expected behavior. In contrast, overfitted or badly optimized parameter sets result in heavily distorted structures as shown in [Figure 6](#), and molecular dynamics show unphysical fragmentations and other artifacts that render the calculations completely useless.

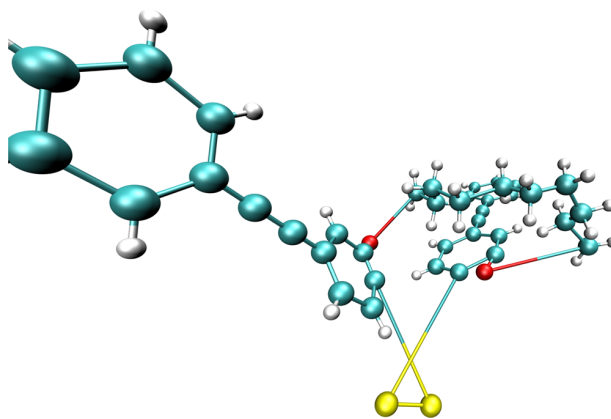


Figure 6. Example for a geometry optimized with an overfitted parameter set. It is obvious that there are problems with the optimal geometry here. As the potential curves in [Figure 5](#) already suggest, there are bonding terms with almost purely repulsive potential curves, which is reflected by the final structure shown here.

When simulations show erroneous behavior for data points outside the reference set, this is a strong indicator for the presence of overfitting. In such a situation the user has two options to resolve the issue, either removing the parameters that are only weakly addressed by the reference set or including additional data in the reference set to stabilize these “wild” parameters.

A common approach to avoid overfitting artifacts, often found in machine learning applications, is an early stopping method.⁴² In this method, a reference set and the validation set are evaluated in parallel. The point where their fitness values start to diverge is assumed to mark the onset of overfitting. This early stopping method was applied in the final phase of the parameter optimization.

3.5. Reference Set Construction and Initial Parameter Optimization. The reference set used in this work was not built from scratch in one step. As mentioned in the theoretical introduction, the REAXFF formalism is too complex to decide a priori which parameters are to be optimized and which reference data would be needed. Therefore, the set introduced here was slowly assembled in a manual feedback loop with the aim of disulfide mechanochemistry in mind and over- and underfitting as limiting factors. Every problem at any stage was closely investigated and eliminated before progressing further. In the initial phase of the optimization, it was not even planned to include SH, SSH, CSC, SCS, or HSSH data in the set, but the interplay between the respective parameters and the parameters of interest made those additions necessary. In the REAXFF formalism, any atom in close proximity to the moiety of interest is treated as partially bound to that group. This leads to situations where parameters that are not well optimized for the problem at hand can interfere with the parameters that are to be optimized. This deforms the potential energy surface in an undesired and unphysical fashion. To repair this kind of problem, the problematic parameters need to be included in the parameter vector subjected to optimization. This in turn requires the reference set to be expanded to get rid of overfitting artifacts that are easily acquired by increasing the number of free parameters.

Parallel to the work on the reference set, the parameter vector subjected to optimization was refined. This refinement entails choice of the parameters and setting their boundaries in the search space.

As a starting point for the parameter optimization, the force field by Mattsson and co-workers⁴³ was chosen. In the original publication, the parameter set contained parameters for carbon, hydrogen, oxygen, nitrogen, and sulfur atoms. Since our structure of interest features no nitrogen atoms, all parameters that apply to nitrogen were removed from the force field file. In a second step, the complexity of the force field was reduced to the necessary level, by removing parameters for all sulfur bond orders higher than one, such that only sigma bonds were allowed.

Choosing the parameters subjected to optimization is crucial for the success of the reparametrization. Starting with 13 optimizeable parameters that are used to describe the SS single bond, the parameter file was grown to the final size of 87 entries. Since the reference set and the parameter set are closely related, it comes at no surprise that the final set contained parameters for the carbon and sulfur atom as well as for CS, SH, SS, CCS, CSC, CSS, HSH, HSS, SCS, CSSC, and XSSX terms. XSSX parameters are used whenever a dihedral contributions for any four-body interaction is calculated that

is not CSSC. The complete parameter file used in the optimization can be found in the [Supporting Information](#).

The choice of the interval boundaries was just as important. Obviously, it determines the actual search space size (within the given dimensionality), which in turn directly translates into smaller or larger search effort, for the same final minimum, as long as it is still enclosed in the boundaries. Hence, the intervals should be as small as possible. On the other hand, small intervals correspond to an a priori bias on the search, dictating that the intervals should be as large as affordable. However, there are further and less obvious considerations to be made: While the more local SOPPE¹⁶ routine as used by van Duin tends to converge to a minimum near the starting parameter set, the evolutionary algorithm is randomly initiated and can jump to any location in the search space due to the (intended) exploratory power of its search operators. If the boundaries of the search space are not carefully chosen, this can lead to various serious problems. The random initialization of the algorithm may generate only individuals with a very bad fitness value, in the worst case even with numerous reference set entries that returned unconverged calculations. This would delay the onset of actual optimization progress and increase the time to convergence significantly. Furthermore, the algorithm might converge to a region in search space that has low deviations from the reference set but is far from any physically reasonable parameter values. Last but not least, relations between parameter values may invert, leading to a situation where, e.g., single bonds are shorter than double bonds. Prior knowledge about reasonable values for every single parameter and the ratios of some of them hence is crucial for successful optimization runs. Otherwise, this knowledge has to be acquired in advance, by additional test calculations.

All modifications at the reference set and parameter vector settings were done in close interplay with the manual validation of the resulting solution vectors as described above.

3.6. Final Parameter Optimization. When the solution vectors found by using the reference set and the parameter vector started to show some convergence toward physically meaningful results, reference set assembly was considered finished. A final optimization phase was then entered, to find the best possible solution vector for this fixed reference set.

Without early stopping, i.e., without considering the validation set, the best individual in the global optimization stage would have reached a final fitness of 7574. Taking the validation set into account, however, indicated that an earlier stop of the global optimization was more appropriate, yielding a reasonable solution vector after 400000 iterations with a fitness value of 12393 and with a better chance for smaller overfitting artifacts. The development of the fitness of the best individual for the reference set and the validation set as a function of the optimization step number is shown in [Figure 7](#).

3.7. Force Field Performance. There currently are two parametrizations for plain REAXFF in the LAMMPS^{44,45} potential library that can describe systems containing carbon, hydrogen, oxygen, and sulfur atoms. One is the parametrization by Singh et al. that was optimized for the simulation of fluorographenes,⁴⁶ the other is the force field by Mattsson and co-workers.⁴³ Both parameter sets produce an error sum of well over 240000 when used on the reference set unaltered. By applying the modifications mentioned in [section 3.5](#) to Mattssons parameter set, the objective value was reduced to 188000. The deviations in energy from the reference surface ranged from about -125 kcal/mol to +40 kcal/mol. By our

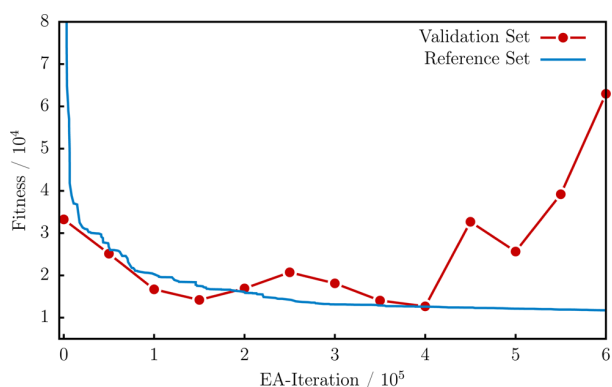


Figure 7. Evolution of the fitness value of the best individual in the pool for the reference set (blue) and the validation set (red). The validation set was evaluated every 50000 iterations.

optimization, the fitness was then improved to a final best value of 12393. Local relaxation with a greedy hillclimb algorithm yielded better fitness values, but at the same time the fitness of the validation set worsened significantly; therefore, the solution vector from the evolutionary algorithm was chosen as the final solution. This solution will be dubbed Mue2016 henceforth. The maximum deviation in energy of this final parameter set was the underestimation of the CS bond dissociation energy in DMTE by 16 kcal/mol. This deviation is depicted in panel (b) of Figure 8.

The potential energy curves in Figure 8 clearly show that the overall qualitative shape of the potential energy surface for various disulfide systems could be greatly improved, compared to the previously existing FFs. This holds true for coordinates

included in the reference set, which are shown in the panels (c)-(d), as well as for data outside the reference set, as the DPDS dissociation curve in panel (a). DPDS has proven to be a good model system for DSM-C and DSM-PEG and is therefore an interesting test case. While minor flaws in the potential curves, e.g., the bond dissociation energy (BDE) and slope of the potential, still remain, the overall quality of the PES was much improved compared to Mattssons parameter set (which was not intended for disulfide mechanochemistry and did not include CASPT2 reference data, so its worse performance for the present purpose was to be expected). The equilibrium distances and angles can be accurately estimated by using the new parametrization. As could be expected, the curvature and slope of the potentials in the panels c-d compare now much better to the CASPT2 results since they were included in the reference set. This trend is not as pronounced for the DPDS dissociation in panel (a), but even for this system that was not included in the reference, the qualitatively wrong behavior of the old parametrization was corrected to a great extent. The overestimation of the slope in the curves should influence the quantitative outcome of force dependent simulations, but qualitative tendencies in covalent mechanochemistry experiments may now be easily captured in simulations.

To benchmark the capabilities of the solution vector to reproduce molecular geometries, Mue2016 was tested on DSM-C and a representative pool of 158 geometries from the group of sulfide and disulfide compounds, most of which had not been part of the reference or validation sets. The results of those tests are shown in Figure 9. Panel (a) shows the direct comparison of the central region of DSM-C in Figure 1. The reference, which was optimized on the MP2 level, is practically indistinguishable from the LAMMPS-REAXFF/Mue2016 optimized

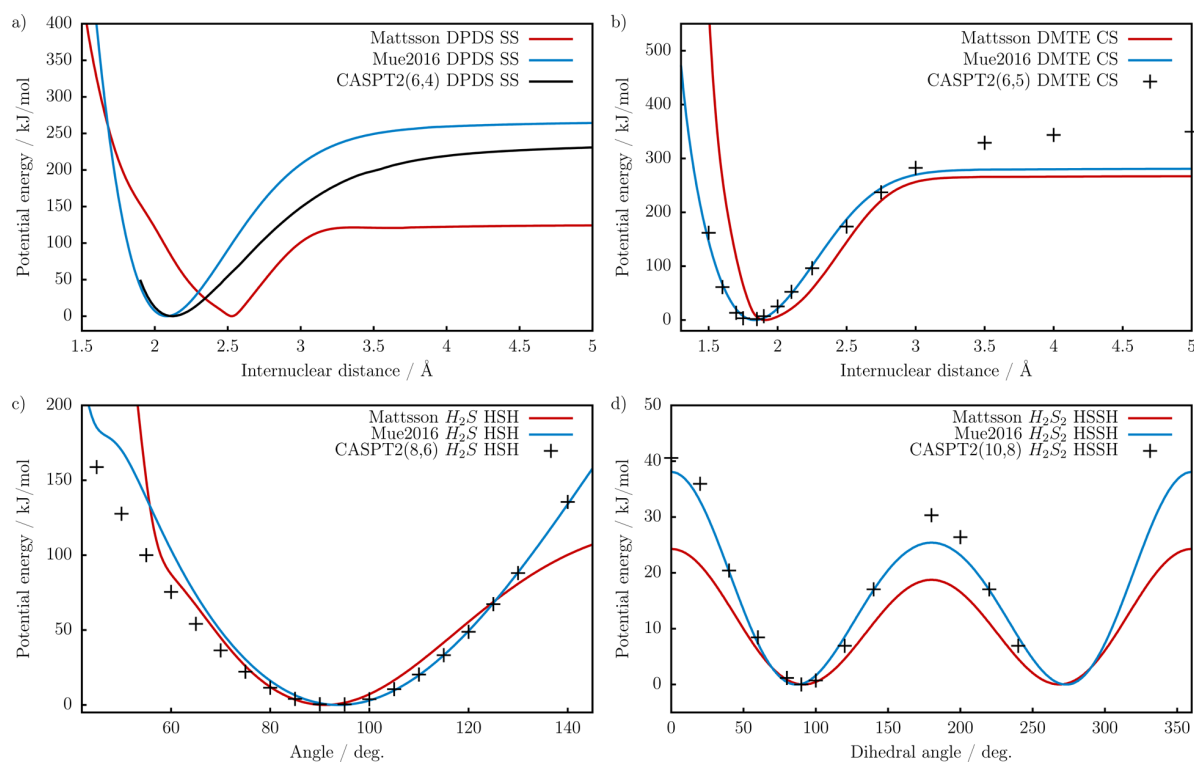


Figure 8. Comparison of various representative potential energy curves, calculated with the new Mue2016 parameter set, with Mattssons force field,⁴³ and with multireference perturbation theory.

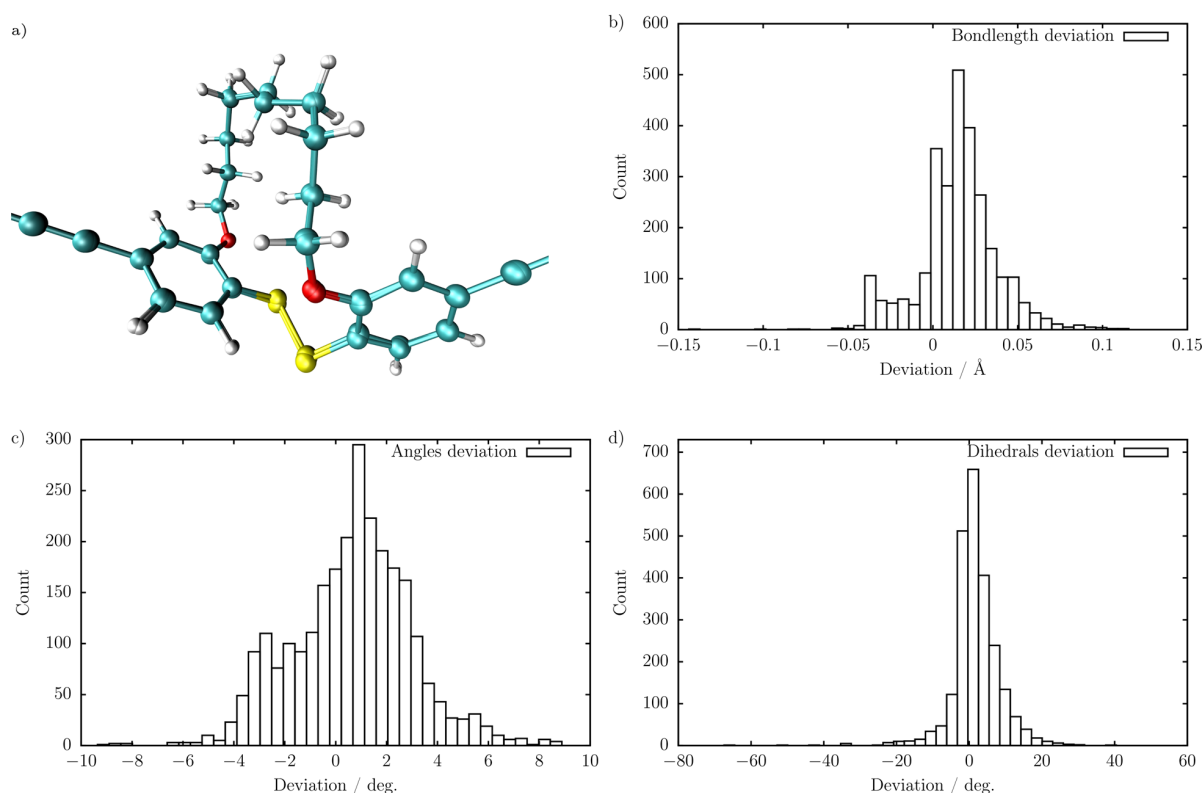


Figure 9. Overview of the geometry optimization benchmark results. The two almost identical geometries superimposed upon each other in the upper left panel are optimized with the RIMP2/cc-pVDZ method and LAMMPS-REAXFF/Mue2016, respectively. The other panels show deviation histograms of internal coordinates for LAMMPS-REAXFF/Mue2016 structures compared to MP2 results.

structure. The remaining three panels show histograms of the deviations of the internal coordinates in this benchmark set. All geometries used to benchmark the solution vector except for DSM-C were optimized with the RIMP2/aug-cc-pVTZ method, as we had done with the reference set structures. The absolute deviations in the coordinates were extracted from the resulting z -matrices of the optimizations and binned to be plotted as histograms. The RMSD values for each coordinate type are 0.0241 Å for the bonds, 2.40 degrees for the angles, and 5.75 degrees for the dihedrals.

Obviously, most deviations are very small except for a few outliers in the dihedral coordinates. One of these outliers is an optimization artifact that can be traced back to small gradients around the minimum of the corresponding geometry, in combination with the not very tight geometry convergence criteria used throughout the global optimization process. These convergence criteria save considerable computer time and are sufficient in almost all cases, with this particular exception. The remaining group of outliers are faulty HSCS dihedrals in the terminal thiol groups of HS-CH₂-SR compounds, which clearly have no relevance for the initial steps of disulfide mechanochemistry. Up to now, these are the only artifacts of this kind that were found in the benchmark data or elsewhere. Hence, these results show that Mue2016 is able to reproduce MP2 geometries with high accuracy.

4. FORCE FIELD APPLICATIONS

To make sure that the solution vector performs well in real-life molecular mechanics settings too, the Mue2016 parameter set was applied in various MD simulations. These simulations were

mainly concerned with molecular behavior under the influence of external mechanical forces as can be exerted by an AFM tip.

All simulations were performed by using the LAMMPS software package with the `reaxc` extension, in the serial version of this code published on February 16th 2016. The external mechanical force was introduced via the `add_force` keyword that allows to add a force vector to the gradient for two anchoring atoms. These anchoring atoms are used to mimic the connection to a surface and to the AFM tip that would be present in an experimental setup.

The calculations were performed with a time step of 0.5 fs at 300 K in the NVT ensemble and a Berendsen thermostat with a damping factor of 100. The random number generator used to generate the initial velocities was seeded with the running number of each respective trajectory to avoid redundant calculations. The simulations themselves included mechanochemistry in vacuo as well as in a toluene solvent box with periodic boundary conditions.

To illustrate the mechanochemical process we target with our simulations, Figure 10 shows how a typical reactive trajectory looks like. To keep this picture as simple as possible, the smaller DSM-C was chosen, and no solvent box is shown. After an initial stretching, the disulfide bond cleaves to a biradical structure, and the safety line unfolds until the maximum length of the system is reached. For events after structure (d) in Figure 10, there are two possibilities. Either the structure remains stable in this final form, or a second dissociation happens somewhere in the molecule, and both isolated parts of the molecule are pulled apart at ever increasing speeds.

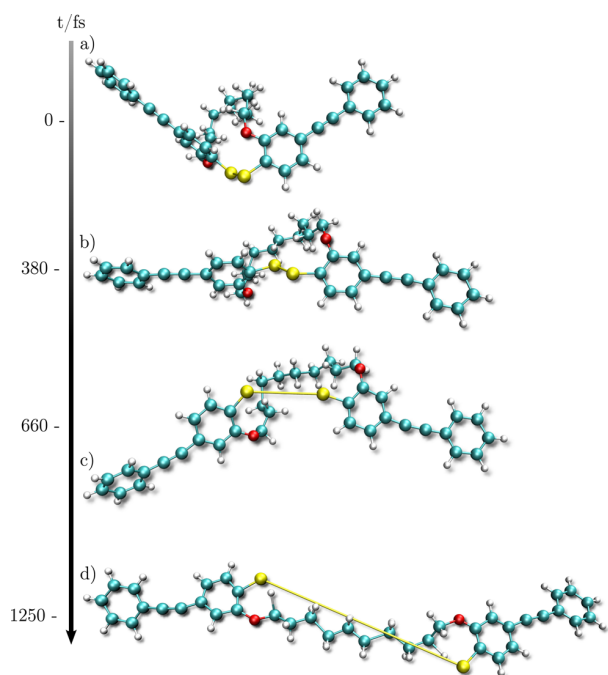


Figure 10. Snapshots from a trajectory with a rupture event of the disulfide bond. a) Undisturbed starting structure at 0 fs propagation time. b) Stretched structure at 380 fs, tension on the disulfide bridge rises. c) Disulfide bond was broken and the safety line starts to unfold at 660 fs. d) The unfolding process is done at 1250 fs. To visually emphasize the big change in sulfur–sulfur distances, a “bond” line is drawn between the two sulfur atoms even for distances far longer than actually bonded distances.

Within the framework of a Bell-type model^{47,48} and the assumption that bond potentials are of a Morse type, it can be shown that the dissociation energy depends exponentially on the applied force. If this dependence is plugged in Arrhenius' equation, a sigmoidal dependence of the rupture rate constant on the applied force results. The central section of this sigmoidal behavior can be fitted by a linear function, to define an unambiguous onset of rupture events in the presence of noise. Furthermore, when simulating force-dependent single molecule experiments, one therefore expects a linear region in the rupture probability distribution for reactive trajectories, within this force interval.

As a first rather crude picture of an experimental AFM setup, trajectories for DSM-PEG were calculated. 35 sets of 250 trajectories each were propagated for 50000 timesteps (25 ps). The forces applied to the ends of the polymer chains were increased by 10 pN in every successive set. The probability for the reaction, which is directly proportional to the reaction rate when looking at trajectories of finite length, was then calculated by dividing the number of reactive trajectories by the total trajectory number at every given force. The resulting probabilities for rupture events, within the simulated time frame of 25 ps, are shown in Figure 11. The red curve shows the force-dependent probabilities for dissociation in the central disulfide moiety, defined as S–S distances greater than 5 Å. The much flatter blue curve shows rupture probabilities of the whole structure, defined as end-to-end distances exceeding 18 nm. This includes events where the polymer or the safety line is breaking and the ends of the polymer are losing touch, irrespective of whether the S–S bond has also been broken or

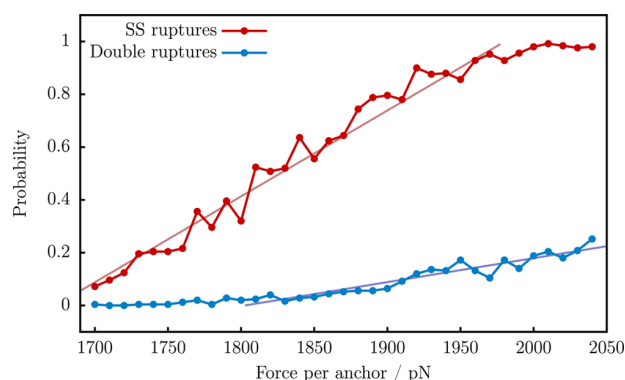


Figure 11. Force dependence of the probability of rupture events in vacuo for a rupture of the SS bond in red and complete rupture in blue. The linear regions are fitted and reveal an onset for rupture events at 1675 pN for the SS coordinate and 1800 pN for the whole structure.

not. These two kinds of events are potentially distinguishable in an AFM experiment, as shown by Schütze.⁴

Fitting the linear region of both probability curves gives an onset for SS dissociation at 1675 pN and at 1800 pN for the rupture elsewhere the structure. According to the linear fit, at 1800 pN the disulfide dissociates with a probability of roughly 0.4. Hence, experimentalists are left with a wide window for reversible mechanochemical switching of these structures. On the other hand, the lifetime of the disulfide bond in vacuo is essentially zero. Every rupture event detected happened on the same time scale as the stretching of the anchoring polymer. This already suggests that there are certain problems with this model, which are discussed in more detail below.

Two problems can be identified here. One is the simple fact that these small test calculations done here cannot capture the enormously longer time scales at which the experiments are conducted. An AFM tip retraction curve is measured on a time scale of seconds. If the region between 1000 pN and 2000 pN pulling force is traversed within 1 s, the mechanophor has roughly 1 ms time per pN to adjust to the external force. This is a time 7 orders of magnitude longer compared to the simulation time of 25 ps used here. Therefore, simulation times have to be much longer to reach into the region of experimental relaxation times. The second problem is momentum gathered by stretching soft degrees of freedom. At the start of any trajectory the soft degrees of freedom stretch out under the influence of the mechanical force. The nuclei gather velocity in the direction of the force, and the momentum gathered rips the weakest bond apart. Collisions with solvent molecules would prevent that from happening since the kinetic energy would be dissipated to the surrounding molecules very quickly and efficiently.

Since both problems were present in the mentioned calculations, the expected decay of the bond lifetimes under the influence of increasing mechanical force could not be observed in vacuo. The significance of these results for experimental applications is therefore rather limited.

To address the problem of artifacts from gathered momentum, DSM-C was solvated in a periodic boundary box with toluene as solvent. The cuboid box with side lengths of $50 \times 25 \times 25 \text{ \AA}^3$ was filled with 161 toluene molecules and equilibrated using the TINKER⁴⁹ software and OPLS-AA⁵⁰ parameters as found in the TINKER force field parameter sets.

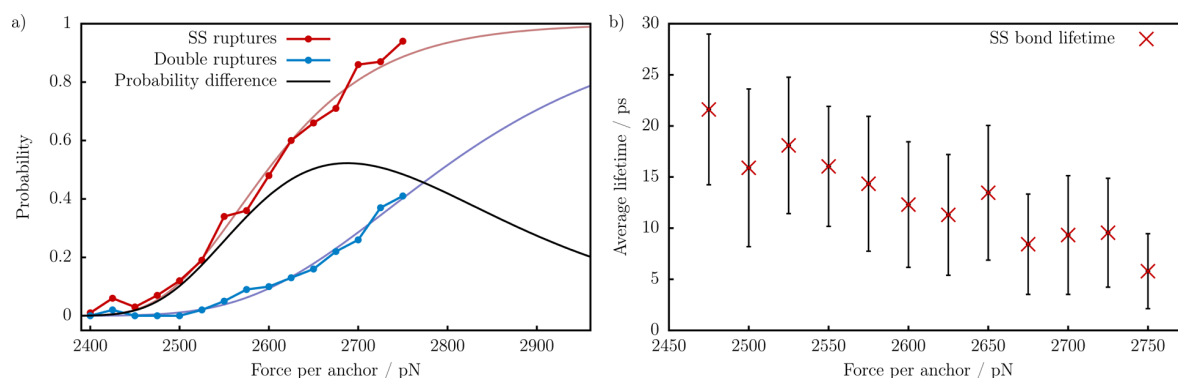


Figure 12. Panel (a) shows the probabilities for dissociation events of the disulfide bond (red) and the structure in total (blue) with respective exponential fits. The difference curve (black) shows the probability to observe an isolated SS dissociation without a following disconnection of the structure anchors. Panel (b) shows the average lifetimes of the disulfide bond in reactive trajectories plotted over the applied mechanical force. Error bars indicate the mean average deviation.

In a second preparation step, the box was equilibrated again at 300 K with periodic boundary conditions, using the LAMMPS implementation of REAXFF with the Mue2016 parameter set. The resulting box was used as input structure for all following molecular dynamics calculations.

After an initial assessment of the force region where rupture events could be observed within 50 ps of simulation time, 15 batches of 100 trajectories each were calculated at 300 K in the NVT ensemble using a Berendsen thermostat with a damping constant of 100. The interval of force applied to the anchoring atoms ranged from 2400 pN to 2750 pN with an increment of 25 pN for every successive batch.

The applied force is about 700 pN higher than the force used in vacuo. Because the cooling of the internal degrees of freedom of the solvated system is present by collisions with the toluene molecules, no momentum can be gathered in stretching processes. This leads to much higher lifetimes of the bonds. To relocate those lifetimes to the used observation window of 50 ps, the applied force has to be increased.

As for the in vacuo simulations above, the probabilities for rupture of the disulfide bridge and of the structure in total were plotted over the applied force (cf. panel (a) of Figure 12). Additionally, double exponential functions of the form $\exp(-\exp(-a(x-x_0)))$, which are expected to represent the probability for rupture events, were fitted to the data. Since reversible mechanical switching requires the structure to keep contact between AFM tip and surface, the probability for a disulfide dissociation without a following disconnection event is of interest. This probability is then expressed by the difference of both probabilities and plotted as the black curve in Figure 12.

We mentioned above that the external mechanical force has a direct influence on bond lifetimes. This dependence could be observed more easily in solution. Panel (b) of Figure 12 shows the average lifetime of the disulfide bond in all reactive trajectories plotted over the external mechanical force. Though there clearly is a correlation between the lifetime and the applied force, the mean average deviations of these averages, given as error bars in the plot, do not allow for further interpretation of the functional form of this relation. Especially in the low force regime the average is calculated from only a few reactive trajectories and is therefore very unreliable.

Although force fields do speed up dynamics calculations by roughly 5 orders of magnitude, compared to direct ab initio dynamics, a significant fraction of the 12 orders of magnitude

time scale discrepancy between simulations and mechanochemical AFM experiments can only be bridged if high-performance supercomputer hardware and efficiently parallelized software is employed additionally. These latter tools were not available to us. Nevertheless, our newly fitted force field also enables longer reactive trajectories on standard computer equipment and standard software (LAMMPS). Hence, as a proof of principle, a small set of long trajectories for DSM-PEG in vacuo has been calculated. Two long runs were started with an external mechanical force of 1950 pN per anchor. The first trajectory was propagated for 100 million timesteps of 0.5 fs at 300 K. To arrive in the microsecond time regime, another calculation ten times as long was started, but the run was terminated after 57 ns when the structure ruptured. Therefore, a third calculation was done with a reduced force of 1670 pN per anchor, for a realistic chance to avoid complete rupture of the structure within 0.5 μ s of simulated time. This latter run finished within 470.8 h of serial execution on an Intel Xeon E5-1620 at 3.6 Ghz. With a total length of 1 billion timesteps of 0.5 fs each, this throughput corresponds to 25.5 ns simulation time per day for DSM-PEG (200 atoms).

Despite their explorative nature, these longer trajectories already allow for two conclusions. Obviously, even on moderate hardware and without parallelization, the REAXFF formalism allows for simulations of a moderately large system on the microsecond time scale, which already approaches total times needed to model AFM experiments with MD methods. Furthermore, the occurrence of reaction events after such a long simulation time indicates that the calculations can indeed capture purely mechanical bond activations on such long time scales. Additionally, these runs show that the reactivity in vacuo does not have to depend on momentum artifacts as described above.

5. CONCLUSION

In this work we have shown that reactive force field MD is a useful and practically accessible tool to simulate mechanochemical events, with a realistic option to address long-time procedures as for example AFM experiments. We have demonstrated this approach with the example of a rather sophisticated disulfide as mechanochemical moiety and with REAXFF as force field. Obviously, other mechanochemical units and other reactive force fields could be targeted as well, using very similar protocols.

Most of the reactive force fields presented in the literature so far have been trained with DFT data as reference. Clearly, this is suboptimal, in particular for covalent bond breaking, in mechanochemistry, or elsewhere: Even on the ground-state Born–Oppenheimer surface of otherwise electronically simple, closed-shell molecules, non-negligible multireference character may develop during bond stretching. However, there is no a priori reason for reactive force fields to fail or only to work less well when fitted to higher-level *ab initio* theory. This is the second major feature of the work presented here: We have employed MP2 and CASPT2 data as reference, since this was indeed necessary to model disulfide dissociation quantitatively correctly. As expected, there was no indication that REAXFF was inadequate for this higher level of reference data. Thus, in effect, the MD simulations shown above are a good approximation to direct CASPT2 dynamics, which will remain impossible for many decades, for systems beyond a few atoms and for times beyond a few dozens of femtoseconds.

Last but not least, the REAXFF fitting presented here is a real-life application example for the nondeterministic, global FF parameter fitting strategy by advanced evolutionary algorithms, which we developed in previous publications.^{23–25} As explained and demonstrated in detail in section 3, the fitting process still requires diligent attention to several crucial details. However, our EA approach turned out to be the vital ingredient to navigate a difficult search space and to avoid overfitting: As we have shown here, it was possible to obtain a system-specific REAXFF force field that performs very well, both when compared to high-level *ab initio* data and when employed in MD studies targeted toward direct simulation of AFM experiments.

■ ASSOCIATED CONTENT

■ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.6b00461.

Optimized parameter values of the final Mue2016 force field, together with all the files and information constituting both the reference set used in the optimization and the validation set used for the early stopping criterion (ZIP)

In-depth discussion and illustration on how to choose boundaries between which parameters are allowed to vary during global optimization of the force field (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hartke@pctc.uni-kiel.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

It is a pleasure for us to thank Johannes Dieterich (Princeton University) for initiating and constantly developing the general global optimization suite OGOLEM, H. Metin Aktulga (Michigan State University) for the REAXFF reimplementation sPuReMD, and Mark Dittner (University of Kiel) for joining sPuReMD and OGOLEM; their help was vital for the present work. Financial support by the German Science Foundation DFG via Collaborative Research Center SFB 677 “Function by Switching” is gratefully acknowledged.

■ REFERENCES

- (1) Binnig, G.; Quate, C. F.; Gerber, C. *Phys. Rev. Lett.* **1986**, *56*, 930.
- (2) Grandbois, M.; Beyer, M. K.; Rief, M.; Clausen-Schaumann, H.; Gaub, H. E. *Science* **1999**, *283*, 1727.
- (3) Fernandez, J. M.; Li, H. *Science* **2004**, *303*, 1674.
- (4) Schütze, D.; Holz, K.; Müller, J.; Beyer, M. K.; Lüning, U.; Hartke, B. *Angew. Chem., Int. Ed.* **2015**, *54*, 2556.
- (5) Beyer, M. K. *J. Chem. Phys.* **2000**, *112*, 7307.
- (6) Ribas-Arino, J.; Shiga, M.; Marx, D. *Angew. Chem.* **2009**, *121*, 4254.
- (7) Wang, J.; Kouznetsova, T. B.; Niu, Z.; Ong, M. T.; Klukovich, H. M.; Rheingold, A. L.; Martinez, T. J.; Craig, S. L. *Nat. Chem.* **2015**, *7*, 323.
- (8) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 5915.
- (9) Liang, T.; Shin, Y. K.; Cheng, Y.; Yilmaz, D. E.; Vishnu, K. G.; Verners, O.; Zou, C.; Phillpot, S. R.; Sinnott, S. B.; van Duin, A. C. T. *Annu. Rev. Mater. Res.* **2013**, *43*, 109.
- (10) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; Verstraelen, T.; Grama, A.; van Duin, A. C. T. *NPJ. Comput. Mater.* **2016**, *2*, 15011.
- (11) Watanabe, T. *J. Comput. Electron.* **2011**, *10*, 2.
- (12) Popelier, P. L. A. *Int. J. Quantum Chem.* **2015**, *115*, 1005.
- (13) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A., III. *J. Phys. Chem. A* **2001**, *105*, 9396.
- (14) Nakano, A.; Kalia, R. K.; Nomura, K.; Sharma, A.; Vashishta, P.; Shimojo, F.; van Duin, A. C. T.; Goddard, W. A.; Biswas, R.; Srivastava, D.; Yang, L. H. *Int. J. High Perf. Comput. Appl.* **2008**, *22*, 113.
- (15) Aktulga, H. M.; Fogarty, J. C.; Pandit, S. A.; Grama, A. Y. *Parallel Comput.* **2012**, *38*, 245.
- (16) van Duin, A. C. T.; Baas, J. M. A.; van de Graaf, B. *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 2881.
- (17) Zhang, B.; van Duin, A. C. T.; Johnson, J. K. *J. Phys. Chem. B* **2014**, *118*, 12008.
- (18) Srinivasan, S. G.; van Duin, A. C. T.; Ganesh, P. *J. Phys. Chem. A* **2015**, *119*, 571.
- (19) Broqvist, P.; Kullgren, J.; Wolf, M. J.; van Duin, A. C. T.; Hermansson, K. *J. Phys. Chem. C* **2015**, *119*, 13598.
- (20) Larentzos, J. P.; Rice, B. M.; Byrd, E. F. C.; Weingarten, N. S.; Lill, J. V. *J. Chem. Theory Comput.* **2015**, *11*, 381.
- (21) Rice, B. M.; Larentzos, J. P.; Byrd, E. F. C.; Weingarten, N. S. *J. Chem. Theory Comput.* **2015**, *11*, 392.
- (22) Iype, E.; Hütter, M.; Nedeá, S. V.; Rindt, C. C. M.; Jansen, A. P. *J. J. Comput. Chem.* **2013**, *34*, 1143.
- (23) Larsson, H. R.; Hartke, B. *Comput. Meth. Mater. Sci.* **2013**, *13*, 120.
- (24) Larsson, H. R.; van Duin, A. C. T.; Hartke, B. *J. Comput. Chem.* **2013**, *34*, 2178.
- (25) Dittner, M.; Müller, J.; Aktulga, H. M.; Hartke, B. *J. Comput. Chem.* **2015**, *36*, 1550.
- (26) Keten, S.; Chou, C.-C.; van Duin, A. C. T.; Buehler, M. J. *J. Mech. Behav. Biomed. Mater.* **2012**, *5*, 32.
- (27) Rahaman, O.; van Duin, A. C. T.; Goddard, W. A., III.; Doren, D. J. *J. Phys. Chem. B* **2011**, *115*, 249.
- (28) Nielson, K. D.; van Duin, A. C. T.; Oxgaard, J.; Deng, W.; Goddard, W. A., III. *J. Phys. Chem. A* **2005**, *109*, 493.
- (29) Rauhut, G.; Hartke, B. *J. Chem. Phys.* **2009**, *131*, 014108.
- (30) Dieterich, J. M.; Hartke, B. *Mol. Phys.* **2010**, *108*, 279.
- (31) <http://www.ogolem.org/> (accessed July 1, 2016).
- (32) Bandow, B.; Hartke, B. *J. Phys. Chem. A* **2006**, *110*, 5809.
- (33) Aktulga, H. M.; Pandit, S. A.; van Duin, A. C. T.; Grama, A. Y. *SIAM J. Sci. Comput.* **2012**, *34*, C1.
- (34) Hartke, B. *J. Comput. Chem.* **1999**, *20*, 1752.
- (35) Shum, L. G. S.; Benson, S. W. *Int. J. Chem. Kinet.* **1983**, *15*, 433.
- (36) Neese, F. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73.
- (37) Neese, F. *J. Comput. Chem.* **2003**, *24*, 1740.
- (38) Kossmann, S.; Neese, F. *J. Chem. Theory Comput.* **2010**, *6*, 2325.

- (39) Karlström, G.; Lindh, R.; Malmqvist, P. Å; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P. O.; Cossi, M.; Schimmelpfennig, B.; Neogrády, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.
- (40) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P. Å; Neogrády, P.; Pedersen, T. B.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.
- (41) Veryazov, V.; Widmark, P. O.; Serrano-Andrés, L.; Lindh, R.; Roos, B. O. *Int. J. Quantum Chem.* **2004**, *100*, 626.
- (42) Behler, J. *Int. J. Quantum Chem.* **2015**, *115*, 1032.
- (43) Mattsson, T. R.; Lane, J. M. D.; Cochrane, K. R.; Desjarlais, M. P.; Thompson, A. P.; Pierce, F.; Grest, G. S. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2010**, *81*, 054103.
- (44) Plimpton, S. J. *Comput. Phys.* **1995**, *117*, 1.
- (45) <http://lammps.sandia.gov/index.html> (accessed July 1, 2016).
- (46) Singh, S. K.; Srinivasan, S. G.; Neek-Amal, M.; Costamagna, S.; van Duin, A. C. T.; Peeters, F. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 104114.
- (47) Bell, G. I. *Science* **1978**, *200*, 618.
- (48) Ribas-Arino, J.; Marx, D. *Chem. Rev.* **2012**, *112*, 5412.
- (49) <http://dasher.wustl.edu/tinker/> (accessed July 1, 2016).
- (50) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

4.4. Additional Information

4.4.1. Concerning Params Files

In an optimization run for REAXFF parameters the `params` input file provides information about which parameters are to be optimized and what are the boundary values for them. In the original format by *van Duin* the parameters are identified by three numbers which act as their coordinates in the `ffield` file. These are followed by three values. The latter two define the upper and lower boundary a parameter is allowed to take during the optimization. The first one is a parameter only important for *van Duin's* SOPPE routine, but was kept in the input to retain compatibility between different implementations. An example of a `params` file can be found in appendix A.4.

The most striking difference between the SOPPE optimization and an evolutionary algorithm strategy is the quasi-local character of SOPPE compared to the true global search of the EA. This has major implications for the choice of parameter boundaries in the search space. While SOPPE tends to sample the objective function close to the starting point of the search, EAs may hop to any point on the objective hypersurface. Multiple preliminary optimizations indicated that the evolutionary algorithm is almost guaranteed to find solution candidates with very low objective values in regions of the search space without any physically meaningful parameter settings when the boundaries are not set up carefully. The computational expense for the optimizations also critically depends on the size of the search space. Any a priori knowledge about the adjustable parameters will increase the quality of the results and decrease the necessary computation time.

The following extensive section will therefore be dedicated to the discussion of various adjustable parameters, their importance for the shape of the PES and reasonable values for their boundaries in the search space. The aim of this section is to provide a comprehensive manual on how to prohibit force field instabilities or other artifacts from occurring during the global optimization. All guidelines discussed below are summarized in tables, which can be found in the appendix.

To keep the section somewhat organized the energy contributions will be discussed one at a time as in the theoretical section on REAXFF potential functions. Some of the potential functions printed there will be reprinted in the following section, this redundancy is meant to prevent excessive scrolling by the reader.

Organisation of `ffield` Files

The collection of empirical parameters for a REAXFF parameterization is saved in the `ffield` file. For an easier understanding of the following paragraphs, the organisation of the `ffield` file is briefly outlined here. It may also be found in the manual of *van Duin's* original REAXFF implementation^[106]. An example of an `ffield` file is found in the appendix.

The `ffield` file is organized in seven sections. These sections read general parameters, atom parameters, bond parameters, off-diagonal parameters, angle parameters torsional parameters and hydrogen bond parameters.

The section of general parameters is a list of 39 parameters which influence multiple parts of the potential irrespective of the atoms involved. Examples are cutoff radii or general overcoordination parameters. To enable transferability of n-body parameters between different parameterization to some extent it is advised to keep the general parameters constant.

The atom parameter section has element symbol identifiers followed by blocks of 32 parameters each organized in four lines. The element symbols are equal to those occurring in the input and output geometry files.

The following five sections are all organized in the same way. Two, three or four leading integers are followed by a block of parameters. The leading integers indentify the elements the n-body parameters refer to. The numbers are according to the occurrence of the elements in the atoms parameter section. Two leading integers refer to bonds and off-diagonal terms. Three atoms are needed for valence angles and hydrogen bonds. Torsional parameters are identified by four atoms.

Single parameters within an `ffield` file are identified by a three-integer address. The first integer can take values from one to seven and specifies which parameters section of the above mentioned is concerned. The integer 2 for example refers to atom parameters. The second number indentifies the parameters block within this section. The numbers 2 1 would address the first block in the atom parameter section, i.e. the first element, usually the carbon atom. The third and last number then specifies the parameter itself. The address 2 1 1 therefore refers to the first parameter of the first block in section two of the `ffield` file which is usually the carbon σ bond length parameter.

Parametrization of Bonding Energies

As outlined in the theoretical section on the REAXFF formalism, the underlying concept to all energy expressions, not only the bonding energy, is the bond order. It is assumed that a first approximation to the bond orders between any two atoms can be calculated by equation 4.2.

$$\begin{aligned} BO'_{ij} &= BO'_{ij}{}^{\sigma} + BO'_{ij}{}^{\pi} + BO'_{ij}{}^{\pi\pi} \\ &= \exp\left(p_{\text{bo}1} \left(\frac{r_{ij}}{r_0^{\sigma}}\right)^{p_{\text{bo}2}}\right) + \exp\left(p_{\text{bo}3} \left(\frac{r_{ij}}{r_0^{\pi}}\right)^{p_{\text{bo}4}}\right) + \exp\left(p_{\text{bo}5} \left(\frac{r_{ij}}{r_0^{\pi\pi}}\right)^{p_{\text{bo}6}}\right) \end{aligned} \quad (4.2)$$

Depending on the allowed bond order of the atom pair considered this equation contains up to nine adjustable parameters. The σ term for single bonds, the π term for a second bond and the $\pi\pi$ term for the third bond are controlled by three parameters each. A distance parameter r_0 and two bond order parameters p_{bo} shape the decay of the bond order with increasing distance r_{ij} .

As this bond order is only a first approximation it is corrected for over- and undercoordination afterwards by five different factors. Three of those five factors are of no concern here, since they are parametrized by parameters of the general parameters block in the force field. As mentioned above it is advised by *van Duin* to keep those constant to retain some transferability of the parameters between different force fields. The latter two correction function are used to soften the effect of over- and undercoordination correction for atoms bearing lone pairs and are given in equations 4.3 and 4.4.

$$f_4(\Delta_i^{\text{boc}}, BO'_{ij}) = \frac{1}{1 + \exp(-p_{\text{boc}3} \cdot (p_{\text{boc}4} \cdot (BO'_{ij})^2 - \Delta_i^{\text{boc}}) + p_{\text{boc}5})} \quad (4.3)$$

$$f_5(\Delta_j^{\text{boc}}, BO'_{ij}) = \frac{1}{1 + \exp(-p_{\text{boc}3} \cdot (p_{\text{boc}4} \cdot (BO'_{ij})^2 - \Delta_j^{\text{boc}}) + p_{\text{boc}5})} \quad (4.4)$$

These two equations are identical except for the overcoordination term Δ^{boc} which is calculated by subtracting the total number of valence electrons including lone pairs from the total bond order of the atom at hand.

$$\Delta_i^{\text{boc}} = -\text{Val}_i^{\text{boc}} + \sum_{j=1}^{N(i)} BO'_{ij} \quad (4.5)$$

In contrast to $p_{\text{boc}1}$ and $p_{\text{boc}2}$ which are general parameters, the bond order correction

4. ReaxFF Parametrization and Disulfide Mechanochemistry

parameters $p_{\text{boc}3}$, $p_{\text{boc}4}$ and $p_{\text{boc}5}$ are adjustable. They are derived from atom parameters in the force field and therefore introduce additional complexity to the optimization problem. The bond order correction parameters are calculated for pairs of atoms as geometrical mean from the atomic parameters (eq. 4.6).

$$\begin{aligned} p_{\text{boc}3,ij} &= \sqrt{bo_{132,i} \cdot bo_{132,j}} \\ p_{\text{boc}4,ij} &= \sqrt{bo_{131,i} \cdot bo_{131,j}} \\ p_{\text{boc}5,ij} &= \sqrt{bo_{133,i} \cdot bo_{133,j}} \end{aligned} \quad (4.6)$$

Understanding exactly how the functions above are working and changing upon adjusting parameters is crucial for the successful global reparametrization of REAXFF force fields to reactions of interest. One important fact which has to be kept in mind above all others is that there is no specific reference data to address bond orders. They are a mathematical construct and gain their physical meaning only in the context of energy expressions like bonding energies. This means the parameters in the equations 4.2 to 4.4 are very prone to fitting artifacts during the global optimization, which is a big problem since they are the very core of the energy expressions and all other terms depend on them.

Bond radii: The bond radii r_0^x seem as if they have a very clear direct interpretation but when looking at them more closely it becomes obvious that r_0^x and equilibrium bond lengths in an optimized molecule are not equal. This is because the parameter r_0^x occurs only as a linear factor in equation 4.2 to scale the coordinate r_{ij} and is not connected to the position of local extrema of the PES at all. This of course is due to the fact that the attractive and repulsive part of the bonding energy are governed by two separate functions. The true equilibrium distance can only be calculated from the sum of the bond energy and the *Van-der-Waals* energy.

The bond order is strictly decreasing in the allowed region of $r_{ij} > 0$. This behaviour can be seen in figure 4.4. The point $r_{ij} = r_0$ does not correspond to a local extremum of the bond order curve. In case of a single bond the bond order has decayed to $\exp(p_{\text{bo}1})$ at this point.

Although this distance parameter is not the same as an equilibrium distance, the importance of these distances may not be discarded as purely parametric numbers. An evolutionary algorithm can easily end up with a solution vector with mixed up the order of the bond types what causes all kinds of problems. The user has to enforce via the

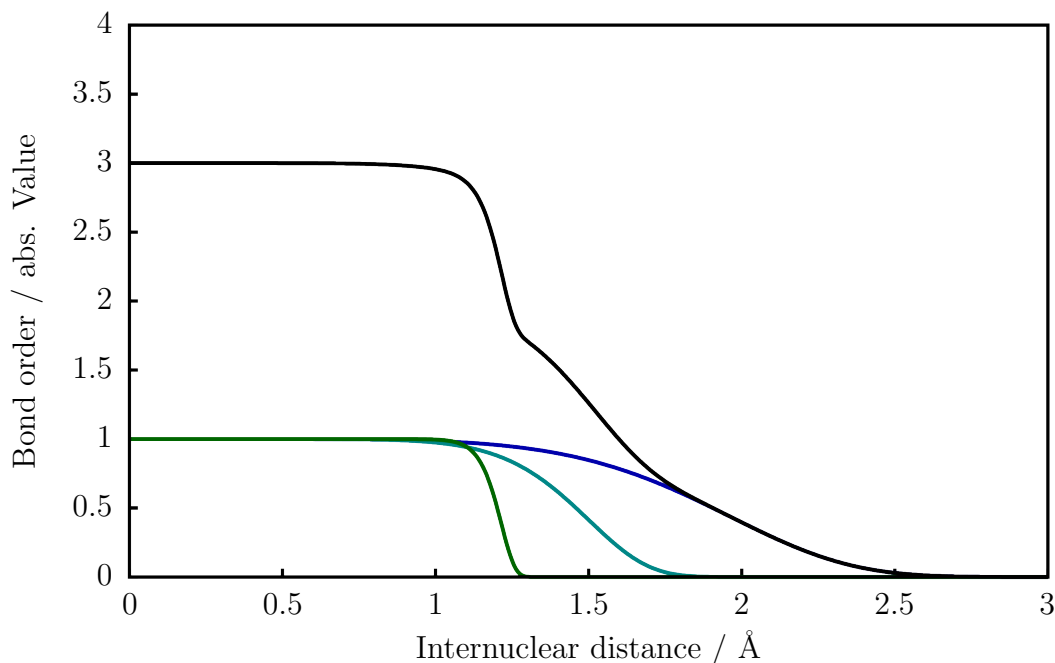


Figure 4.4.: Carbon carbon bond orders using the parameters from *Mattssons* force field as found in the LAMMMPS potential library. The actual parameters used are $r_0^\sigma = 1.3763$, $r_0^\pi = 1.2596$, $r_0^{\pi\pi} = 1.2065$, $p_{bo1} = -0.0994$, $p_{bo2} = 5.9724$, $p_{bo3} = -0.1940$, $p_{bo4} = 8.6733$, $p_{bo5} = -0.7816$ and $p_{bo6} = 28.4167$.

`params` input file that any other solution than $r_0^\sigma > r_0^\pi > r_0^{\pi\pi}$ is impossible, otherwise severe numerical instabilities will occur.

In addition to the atomic covalent radii for different bond orders, the radii may also be specifically parameterized per heteronuclear bond in the off-diagonal section of the force field. The bond radii are then not calculated as mean value from the atomic covalent radii but adjusted for every bond type. This increases the accuracy and flexibility of the force field but complicates the parameterization process. In case of the disulfide parameterization, off-diagonal parameters were used for CS-bonds. For the off-diagonal bond radii the same rules that were discussed before apply.

Undesired bond orders can be deactivated by assigning negative bond radii to either the atom or the off-diagonal parameters. An `if`-condition in the `REAXFF` code checks whether the bond radius parameter is smaller than zero and if this is the case, then no energy contributions are calculated for these bond orders. The optimization problem can be simplified significantly by excluding higher order bonding terms which are not needed for the modeled reactions. For example the disulfide parameterization described

in reference^[3] was not concerned with sulfur bond orders greater than 1, consequentially all higher bonding terms were deactivated. However, with the guidelines for parameter optimizations discussed here and the correct amount of reference data, there is no reason to exclude higher bond orders generally. The reparametrization should be equally successful with all possible bond orders being active.

Bond order parameters: The bond order parameters one through six are used to control the steepness of the bond order decay. Figure 4.4 shows how the bigger values for the bond order parameters influence the corresponding function’s curvature.

Intuitively the order of a σ -bond in ethane should decay much slower with increasing internuclear distances than the $\pi\pi$ -bond of an ethine molecule. This is reflected in the fact that the higher-order parameters have larger numerical values. Similar to the situation described above cases where higher order parameters have smaller absolute values than their lesser order counterparts must be avoided.

The size of the search space can be reduced by looking at reasonable values for the bond order parameters. Obviously the bond order parameters one, three and five have to be negative numbers, otherwise the bond order would grow exponentially with increasing distance. In the paragraph on bond radii it was already mentioned that the bond order decays to $\exp(p_{bo1})$ at $r_{ij} = r_0^\sigma$. To retain at least some relation between those radii and equilibrium distances, and furthermore between dissociation energy parameters and calculated BDEs, the bond order should not be much smaller than 0.5 at $r_{ij} = r_0^\sigma$. This means the upper boundary for bond order parameters one, three and five is roughly -0.7 .

Bond order parameters two, four and six must be positive numbers and larger than 2. with powers less than 2, or the sigmoidal character of the bond order curve would start to vanish and become a simple exponential decay at values of 1. Typical values are around 5 for σ bonds, 5-10 for π bonds and greater than 15 for $\pi\pi$ bonds.

Over- and Undercoordination parameters: The parameters p_{boc3} , p_{boc4} and p_{boc5} are the geometric mean value of the three different atomic parameters bo_{131} , bo_{132} and bo_{133} . They are calculated for each possible pair of atoms according to equation 4.6. The effect on the PES of these parameters is not as easily anticipated as for the previously discussed cases. This is because the influence is not as fundamental as that of the bond order parameters and the functions are more convoluted.

Since the bond order correction parameters can not be interpreted physically very

4. ReaxFF Parametrization and Disulfide Mechanochemistry

well and the objective function for them becomes to complex and highdimensional to be mapped, other approaches were needed to line out their boundaries in the search space. In this case this implied looking at published parameter sets and finding out which parameter setting can produce stable results.

The parameters bo_{131} , bo_{132} and bo_{133} assume a wide range of values from the earliest published force fields to recent ones in the LAMMPS potential library. Some systematic trends can be identified for atoms like Oxygen and Hydrogen, for other atom types the behaviour is more erratic. Values between 2 and 45 are found for bo_{131} and bo_{132} , while bo_{133} covers the region from 0 to 15.

Such “soft” parameters are difficult for the evolutionary algorithm. They vary over a wide range of values while having little impact on the PES and subsequently on the objective function value. Parameters like these are very likely to cause overfitting artifacts. They furthermore contribute to the domino-convergence problem mentioned in section 2.3.1. To overcome those problems, preliminary optimizations with a large searchspace (e.g. 0 – 50 for all three parameters) were done. The solution vectors were thoroughly checked for unphysical behaviour and the search space reduced to regions with strong localisation of good solution candidates for the following optimization runs. Examples for such localisations of parameter values over all solution candidates in a pool can be found in the supporting information of our publication^[3].

Bonding energy parameters: The bonding energy E_{bond} can be calculated using the corrected bond orders^[12]. Equation 4.7 is employed to compute all pairwise bonded interactions for atom pairs within the cutoff range. It introduces three to five adjustable parameters to the equations, depending on the bond order allowed for the atom pair.

$$E_{\text{bond}} = -D_e^\sigma \cdot BO_{ij}^\sigma \cdot \exp(p_{\text{be}1}(1 - (BO_{ij}^\sigma)^{p_{\text{be}2}})) - D_e^\pi \cdot BO_{ij}^\pi - D_e^{\pi\pi} \cdot BO_{ij}^{\pi\pi} \quad (4.7)$$

The dissociation energy parameters D_e^x are used to adjust the depth of the potential minimum along the bonding coordinate. It should be mentioned here that, similar to the bond radius parameters, the parameters D_e^x are related to the BDE of a bond, but they aren’t equal or even lineary dependent on each other. This is due to the fact that the bond orders BO_{ij}^x are not unity at the equilibrium distance of the bond. As mentioned above, the bond order roughly decays to 0.6 – 0.8 at this point. Therefore the parameters D_e^x have to be chosen to be substantially larger than the expected BDEs of the corresponding bonds. The boundaries for these parameters in the evolutionary

4. ReaxFF Parametrization and Disulfide Mechanochemistry

optimization have to reflect this behaviour. An easy way to set them up is using the actual BDE against which the parameters will be optimized and scale it by appropriate factors. A possible approach is shown in equation 4.8. When no reliable BDEs are available it is also possible to estimate the BDE from similar molecular structures and widen the ranges somewhat.

$$D_{e,\min} = 0.85 \cdot \frac{\text{BDE}}{0.7} ; D_{e,\max} = 1.15 \cdot \frac{\text{BDE}}{0.7} \quad (4.8)$$

One not so obvious complication which might occur during the evolutionary optimization is that the dissociation energy parameters D_e^x tend towards high values when given the possibility, in fact much to high values to be physically meaningful anymore. Solution vectors with D_e^x three or four times the BDE can easily be generated. Such a harsh overestimation of D_e^x generates additional flexibility for harder parameters for example the bond order parameters, especially the six bond order parameters $p_{\text{bo}x}$. This contributes to the overestimation of BDEs discussed before.

As stated in *van Duins* original publication on REAXFF from 2001^[12], the repulsion energy of bonds is governed by the *van-der-Waals* energy term. This contrasts most traditional force field implementations which feature local extrema along the bonding potential energy curve at the equilibrium distance. Furthermore it is the reason why the parameters r_0^x are not the equilibrium distance of bonds in the REAXFF formalism.

To ensure that no artificial minima occur in E_{bond} , the bond energy parameters $p_{\text{be}1}$ and $p_{\text{be}2}$ must be chosen accordingly. As BO_{ij}^σ is strictly decreasing with r_{ij} and has limiting values of 1 for $r_{ij} = 0$ and 0 for $r_{ij} \rightarrow \infty$, it is sufficient to search for extrema of E_{bond} in BO_{ij}^σ for $0 < BO_{ij}^\sigma < 1$. Doing the calculus for the bonding energy term (eq. 4.7) yields an expression for extrema in BO_{ij}^σ given in equation 4.9.

$$BO_{ij}^\sigma = \left(\frac{1}{p_{\text{be}1} \cdot p_{\text{be}2}} \right)^{\frac{1}{p_{\text{be}2}}} \quad (4.9)$$

This condition can only have real valued solutions for $\text{sgn}(p_{\text{be}1}) = \text{sgn}(p_{\text{be}2})$. Since no physically meaningful energy curves can be obtained for $p_{\text{be}2} \leq 0$, artifacts may only originate from regions of the searchspace where $p_{\text{be}1} > 0$ and $p_{\text{be}2} > 0$. With the constraint that $0 < BO_{ij}^\sigma < 1$ extrema only occur when $p_{\text{be}2} > 1/p_{\text{be}1}$.

Taking all published parameter sets into account which are currently included in the LAMMPS potential library reveals ranges of $-1.2 \leq p_{\text{be}1} \leq 1.0$ and roughly $0 < p_{\text{be}2} \leq 20$ for the bond energy parameters. It is therefore quite possible that the above mentioned situation can occur during the global optimization. It is therefore necessary

to have a keen eye on the parameter settings after the optimization when the boundaries can not be setup to prevent extrema from occurring directly from the start.

Parametrization of Valence Angle Energies:

The valence angle energy contributions of atoms i,j and k for a system are calculated according to equation 4.10.

$$E_{\text{val}} = p_{\text{val1}} \cdot f_7(BO_{ij}) \cdot f_7(BO_{jk}) \cdot f_8(\Delta_j^{\text{boc}}) (1 - \exp(-p_{\text{val2}}(\Theta_0(BO) - \Theta_{ijk})^2)) \quad (4.10)$$

The functions $f_7(BO_{ij})$ and $f_8(\Delta_j^{\text{boc}})$ are used to make sure that no angle energy contributions are calculated for unbound atoms. Their functional form is given in equations 4.11 and 4.12

$$f_7(BO_{ij}) = 1 - \exp(-p_{\text{val3}}(BO_{ij})^{p_{\text{val4}}}) \quad (4.11)$$

$$f_8(\Delta_j^{\text{boc}}) = p_{\text{val5}} - (p_{\text{val5}} - 1) \frac{2 + \exp(p_{\text{val6}} \cdot \Delta_j^{\text{boc}})}{1 + \exp(p_{\text{val6}} \cdot \Delta_j^{\text{boc}}) + \exp(-p_{\text{val7}} \cdot \Delta_j^{\text{boc}})} \quad (4.12)$$

The remaining function $\Theta_0(BO)$ in equation 4.13 is used to calculate the equilibrium angle of an atom triple. Since there are no atom types as in most non reactive formalisms it must be ensured that the different coordinations for the different possible bonding motifs are accounted for.

$$\Theta_0(BO) = \pi - \Theta_{0,0} \cdot (1 - \exp(-p_{\text{val10}} \cdot (2 - SBO2))) \quad (4.13)$$

The function $SBO2$ returns a value between 0 and 2 depending on the bond order BO_{ij} , BO_{ij}^π and $BO_{ij}^{\pi\pi}$. The parameter p_{val10} is a general parameter and therefore considered a global constant. hence the only remaining adjustable parameter in this function is $\Theta_{0,0}$.

The total number of adjustable parameters to control the angle terms is therefore seven. As done for the bonding energy above, their effect on the PES and reasonable parameter boundaries in the search space will be discussed in detail.

Angle bond order parameters: The parameters used in the equations for $f_7(BO_{ij})$ (eq. 4.11) and $f_8(\Delta_j^{\text{boc}})$ (eq. 4.12) again have no clear physical interpretation.

4. ReaxFF Parametrization and Disulfide Mechanochemistry

As $f_7(BO_{ij})$ obviously goes to zero when the bond order vanishes, the global optimization needs to make sure $p_{\text{val}3}$ and $p_{\text{val}4}$ are chosen to make $f_7(BO_{ij})$ unity when the bond order goes to one. Typical values for $p_{\text{val}3}$ range from 2 to 4. The parameter $p_{\text{val}4}$ takes values between 1 and 4 in published parameterizations. Within these ranges the function $f_7(BO_{ij})$ does not change drastically and they can therefore be safely used for global optimization.

The parameters $p_{\text{val}5}$, $p_{\text{val}6}$ and $p_{\text{val}7}$ in $f_8(\Delta_j^{\text{boc}})$ (eq. 4.12) are controlling the effect of overcoordination on the valence energy. As it was the case for $p_{\text{val}10}$ above, $p_{\text{val}6}$ is part of the general parameters and therefore assumed to be constant for the parameterization. The remaining two parameters $p_{\text{val}5}$ and $p_{\text{val}7}$ again lack clear physical interpretations and thus good estimates for their values. Published parameterization sets show values for $p_{\text{val}5}$ between 2 and 4, $p_{\text{val}7}$ is distributed between 0 and 3. Plotting $f_8(\Delta_j^{\text{boc}})$ in these parameter ranges reveals no problematic behaviour for evolutionary optimization.

All four parameters discussed here do not alter the valence angle energy drastically when changed by substantial amounts, they are therefore all prone to underfitting. Again the domino-convergence is a problem with these soft parameters. Countermeasures like regularization methods or shrinking the search space through preliminary runs have to be taken.

Equilibrium angle: The equilibrium angle for an atom triple is calculated using a function (eq. 4.13) which depends on the parameter $\Theta_{0,0}$ and the σ , π and $\pi\pi$ bond orders of the central atom. Taking a carbon-centered species as an example leads to various bonding motifs with varying geometries. Tetrahedral sp^3 species have an equilibrium angle of roughly 110° , planar alkenes are sp^2 hybridized and show angles of 120° , finally alkynes or species with cumulated double bonds show sp hybridizations and are linear. These possibilities need all to be captured by one parameter $\Theta_{0,0}$.

This is done by the latter expression of equation 4.13. The term $(1 - \exp(-p_{\text{val}10} \cdot (2 - SBO2)))$ goes to zero when the bond orders of π and $\pi\pi$ bond sum up to 1.9 or above, in this case the equilibrium angle will always be 180° . When only σ bonds are present the term value tends towards 1, therefore the equilibrium angle becomes $\pi - \Theta_{0,0}$. In case of alkenes, the behaviour of $SBO2$ is not very obvious, but the equilibrium angle will roughly be $\pi - \Theta_{0,0} \cdot 0.85$. From the argument above it follows that 70° is a reasonable choice for $\Theta_{0,0}$ of the carbon-centered angles and other angles which show similar geometric behaviour. Disturbed angles that deviate just slightly from tetrahedral or trigonal planar geometries can be easily captured with a search space of 60° to 80°

for $\Theta_{0,0}$. When more unusual geometries or higher valencies are considered the user may want to increase the boundaries either to 50° or to 90° , which would be enough to consider for example octahedral structures.

Valence energy: The valence energy term E_{val} (eq. 4.10) is an inverted gaussian type function centered at the equilibrium angle Θ_0 . The parameters p_{val1} and p_{val2} are used to control the depth of the well and the width of the bell curve, respectively.

Currently there is no property available to address these parameters directly by reference data. Force constants or full Hessians could be used to stabilize such force constant type parameters better, but they are not implemented now.

It is therefore more difficult to give practical advice for the choice of both parameters since it is not usual to think in asymptotic energies for angles. The force constant in the vicinity of the equilibrium angle derived from a Taylor expansion is $2 \cdot p_{\text{val1}} \cdot p_{\text{val2}}$. Due to the form of equation 4.11, it seems reasonable to choose p_{val1} to be two times the height of the barrier between two minima along the angle coordinate. For example for water this would mean stretching the HOH-angle from the equilibrium geometry to 180° , doing this results in a barrier of about 7 kcal/mol. A reasonable value for p_{val1} is therefore in the range of 15 kcal/mol. The parameter tolerates a relatively large interval for the evolutionary search algorithm, if overfitting issues are captured with regularization techniques, the algorithm can easily deal with $\pm 30\%$ of the estimated value.

For the second parameter p_{val2} , a wide range of values is acceptable. Values as low as 0.5 can be found in published force fields and just at the value of 10.0 the curvature of the potential starts getting so high that further increase of p_{val2} has no physical justification. If there is a reliable estimate for the harmonic force constant of a valence angle term p_{val2} can be chosen accordingly with respect to p_{val1} and with a narrow search region of ± 2 . If no such information is available the whole range from 0.5 to 10.0 may be searched for solutions with the necessary precautions concerning overfitting.

Parameterization of Dihedral Energies

The torsional energy for four atom i,j,k and l is calculated by equation 4.14.

4. ReaxFF Parametrization and Disulfide Mechanochemistry

$$\begin{aligned}
 E_{\text{tors}} &= f_{10}(BO_{ij}, BO_{jk}, BO_{kl}) \cdot \sin(\theta_{ijk}) \cdot \sin(\theta_{jkl}) \cdot E'_{\text{tors}} \\
 E'_{\text{tors}} &= \frac{1}{2} (V_1(1 + \cos(\omega_{ijkl})) + F_{BO} \cdot V_2(1 - \cos(2\omega_{ijkl})) + V_3(1 + \cos(3\omega_{ijkl})) \quad (4.14) \\
 F_{BO} &= \exp \left(p_{\text{tor1}} (BO_{jk}^\pi - 1 + f_{11}(\Delta_j^{\text{boc}}, \Delta_k^{\text{boc}}))^2 \right)
 \end{aligned}$$

The four adjustable parameters in this expression are the rotational barrier heights V_1 , V_2 and V_3 and the torsional parameter p_{tor1} . The bond order dependent term which scales the second rotational barrier decays with increasing double bond character on the central two atoms of the dihedral to avoid high repulsive contributions for the planar sp^2 geometries. In single bonded cases the term is near unity, for triple bonded sp geometries the $\sin(\theta_{ijk})$ and $\sin(\theta_{jkl})$ terms are taking care of unphysical contributions as they vanish for linear geometries.

The parameter p_{tor1} has no profound impact on the shape of F_{BO} and may safely be optimized to values between -1.0 and -10.0 . Again, as this parameter is not very influential, overfitting or underfitting needs to be avoided.

The rotational barriers V_1 , V_2 and V_3 can take almost any value from -100.0 to 100.0 . The total rotational potential is then a linear combination of the functions $(1 + \cos(\omega_{ijkl}))$, $(1 - \cos(2\omega_{ijkl}))$ and $(1 + \cos(3\omega_{ijkl}))$ with the weighting coefficients V_1 , V_2 and V_3 . Reasonable starting values for the barrier heights in the global optimization may be obtained by fitting a simplified torsional potential (eq. 4.15) to reference dihedral scans if available.

$$E_{t,\text{simple}} = 0.4 (V_1(1 + \cos(\omega_{ijkl})) + V_2(1 - \cos(2\omega_{ijkl})) + V_3(1 + \cos(3\omega_{ijkl})) \quad (4.15)$$

Having good starting values and being able to set narrow search space boundaries around them is crucial for a successful optimization. Due to the number of dihedral contributions which are calculated in bigger systems the total torsional energy can make up a substantial amount of the total energy. At the same time the number of items in the reference set designated to the parameterization of one specific torsional energy profile is usually rather limited. It is therefore very likely that harsh overfitting artifacts are occurring which can be traced back to misaligned rotational barriers. The overfitting artifacts in the rotational barriers then make up for errors in two- or three-body terms. If it is not possible to get reliable estimates for the rotational barriers prior to the global

optimization, preliminary EA runs need to be performed to narrow down the range of the rotational barriers by comparing the REAXFF output to rotational profiles at the ab-initio level.

It should also be noted that the second-order barrier decays rapidly with the π bond order. During the preliminary sampling of the search space for initial guesses for the rotational profile it is therefore necessary to be aware of the bond orders occurring for the torsional PES data in the reference set. Initial parameterization with high barriers V_2 will lead to problems when only single-bonded torsional terms are in the reference set.

Parameterization of Non-Bonded Interactions

The final part of this section is concerned with the non-bonded interactions, namely *van-der-Waals* interactions (eq. 4.16) and *Coulomb* interactions (eq. 4.17).

$$E_{\text{vdw}} = \text{Tap}(r_{ij}) \cdot D_{ij} \cdot \left[\exp \left(\alpha_{ij} \left(1 - \frac{f_{13}(r_{ij})}{r_{\text{vdw}}} \right) \right) - 2 \exp \left(\frac{1}{2} \alpha_{ij} \left(1 - \frac{f_{13}(r_{ij})}{r_{\text{vdw}}} \right) \right) \right]$$

$$f_{13}(r_{ij}) = \left[r_{ij}^{p_{\text{vdw}1}} + \left(\frac{1}{\gamma_{\text{vdw}}} \right)^{p_{\text{vdw}1}} \right]^{\frac{1}{p_{\text{vdw}1}}}$$
(4.16)

$$E_{\text{Coul}} = \text{Tap}(r_{ij}) \cdot C \cdot \frac{q_i q_j}{\left[r_{ij}^3 + \left(\frac{1}{\gamma_{ij}} \right)^3 \right]^{\frac{1}{3}}}$$
(4.17)

The Taper function $\text{Tap}(r_{ij})$ is used to make sure that non-bonded interactions vanish as the distance between the atoms approaches the cutoff radius to avoid discontinuities along the potential curve.

The *van-der-Waals* interactions are parameterized by four adjustable parameters. The parameter D_{vdw} is the depth of the *van-der-Waals* potential well. There are two possibilities to set the parameter, either it is chosen for every atom and then calculated as the geometric mean value for the atom pairs or it is set as an off-diagonal parameter for a specific atom pair directly. Since D_{vdw} has a trivial physical interpretation it is easy to find reasonable values and boundaries for the search. In most application cases 0.0 to 1.0 is a good search region.

The radius r_{vdw} is half the equilibrium distance of the *van-der-Waals* potential. Again it can be set up as an atomic parameter or as an off-diagonal parameter for a specific

atom pair. When atomic parameters are used, the equilibrium distance is two times the geometric mean of both radii, in case of a bond specific parameter it is just the parameter multiplied by two. This is one of the few parameters with an unaltered direct interpretation, it is therefore quite easy to find reasonable starting values for the *van-der-Waals* radius and narrow boundaries in the search space.

The parameter α_{ij} is the force constant of the *van-der-Waals* potential and controls the width of the potential well. It can safely be chosen to be in the range from 5 to 15.

The fourth parameter γ_{ij} damps the coordinate when the distance between two atoms comes close to zero and avoids therefore excessively high repulsion energies. As r_{ij} approaches zero the coordinate function $f_{13}(r_{ij})$ goes to $1/\gamma_{ij}$. Values greater than 2 ensure that the effect on the potential curve is vanishing at $r_{ij} = r_{eq}$ and a maximal value of 10 still yields enough shielding.

Recommendations for the QEq parameters are not that easily given as misaligned parameters can have drastic effects on the force field stability. For almost all elements EEM parameters χ and η are given in the literature. They deviate sometimes more sometimes less between different parameterizations and implementations of EEM type methods. In general trends can be found for one element. Literature values should be used as starting points and varied only in a narrow region around the initial value during the optimization. The shielding parameter is also advised to be varied only by small amounts and not reach values below 0.7 and above 1.2 at all.

In general when starting from an already well tested parameterization there is usually no need to reoptimize the QEq parameters. For this reason, and because the parameters caused more issues than solved them, no QEq parameters were optimized in our publication^[3].

4.4.2. Strained Molecular Dynamics

The resulting parameter set discussed in our publication^[3] was able to produce good results for the kinetics and dynamics of strained polymeres featuring a disulfide moiety in vacuo and in solution. However, this section will be dedicated to discuss another difficulty in the empirical potential description of strained molecular dynamics.

The example quantity will again be the change in length of the structure upon rupture of the predetermined molecular breaking point. The elongations have been evaluated by measuring the anchor to anchor distance of the mechanophores used in the publication. The results are compiled in tables 4.1 and 4.2.

These values may be compared with those from the chapter on triazole. Here the

Table 4.1.: Structural elongations of the disulfide mechanophore in vacuo. The lengths of the structures vary by up to 0.5 Å for different trajectories with the same straining force.

Force / nN	Closed length / Å	Open length / Å	Elongation / Å
3.5	91.5	101.7	10.2
3.6	91.4	103.2	11.8
3.7	91.5	103.9	12.4
3.8	91.3	103.4	12.1
3.9	91.6	103.5	11.9

Table 4.2.: Structural elongations of the disulfide mechanophore in solution. The lengths of the structures vary by up to 0.3 Å for different trajectories with the same straining force.

Force / nN	Closed length / Å	Open length / Å	Elongation / Å
4.9	25.7	36.6	11.0
5.0	25.8	36.7	10.9
5.1	25.8	36.7	10.9
5.2	25.9	36.7	10.8
5.3	25.8	36.8	11.0
5.4	25.8	36.8	10.9
5.5	26.0	36.9	10.9

same safety line was used as for the triazole, but compared to the 1,2,3-triazole DFT calculations show that the disulfide system is shorter. At the low rupture force of 1.11 nN measured experimentally for the 1,2,3-triazole^[2] the change in length should already be about 14 Å and getting larger with increased straining forces. The data shows that the length is almost constant and substantially smaller than the expected result.

The substantial underestimation of the elongation in the disulfide case can be traced to the potential curves of bonds which have not been reparameterized. The CC-bonds and CO-bonds still use the parameters published by *Mattsson*. These potentials feature too high dissociation barriers and erroneous gradients in the region of steepest increase. These stiffer potentials lead to an underestimation of the structure length under external strain.

Depending on the application case the reparameterization must include these parameters and shape the potential correctly. If only the dynamics of the central disulfide moiety is of interest, the optimization of the small parameter set with 87 adjustable

parameters suffices. For the investigation of strained geometries on a larger scale, all relevant parameters need to be taken into account.

4.5. Conclusion

The application of reactive molecular dynamics to covalent mechanochemistry or more specifically the parameterization of empirical potentials for certain mechanophores was not a trivial task. Many hidden obstacles that occurred throughout the project have been addressed and discussed above. With the knowledge about the intricacies of the REAXFF functions and the interactions of the parameters with the EA which was compiled in this thesis it will be easier in the future to parameterize force fields for specific applications. To further increase the performance of OGOLEM for the global optimization of REAXFF parameters the code could be extended.

A big improvement would be the addition of regularization functions. In our 2016 publication the problem of overfitting was discussed at length^[3]. The early stopping method described there was effective but inconvenient to use. The second reference set that was used to assess the overfitting was supplied manually after the calculation was done. Results had to be extracted in a quite convoluted fashion. This whole procedure is a waste of computational resources and work time of the user. Including the early stopping directly into the OGOLEM code would have several advantages. The second reference set can be supplied in a convenient way or generated during the calculation as subset of the main reference set. When the objective value of the comparison set is calculated during the optimization, the necessity to save the large intermediate pool files vanishes. It could even be possible to define a solid termination criterion as there is no need for further optimization beyond the onset of overfitting.

Another direct increase of performance can be obtained when the mandatory restrictions on parameter values that were mentioned in section 4.4.1. Immediately after the generation step of solution candidates they may be checked for errors (e.g. $r_0^\pi > r_0^\sigma$) and erroneous candidates should be discarded. This would mean an on-the-fly restriction of the search space while the maximal flexibility of the parameters is retained.

5. Parameterization of a 1,2,3-triazole force field

In the scope of an advanced practical course under my supervision, *Julien Steffen* assembled a reference set and optimized parameters for the 1,2,3-triazole system introduced in chapter 3^[4]. In the following paragraphs a summary of *Steffen's* project will be given before my follow-up work is discussed.

The idea for the project originated from the questions regarding the 1,2,3-triazole SMFS experiments that were discussed at length in the previous chapter. A dynamic description of the force induced AAC promised to give further mechanistic insight into the experimental data and maybe even yield elongations that compare better with the experimental findings. The parameterization of the disulfide system was already ongoing and therefore a sufficient amount of know-how could be put to use to tackle the 1,2,3-triazole parameterization in REAXFF. It should be mentioned that the project was started well before it was conclusively shown that the reversed AAC does not occur in the AFM experiments.

5.1. Work by Julien Steffen^[4]

5.1.1. The Parameters

The optimization was done with two different parameterizations as starting points, the forcefield by *Mattsson* and coworkers^[107] as well as the parameters set by *Rahaman et al*^[108]. The `params` files used for the optimization was obtained with the automated procedure described before. During the procedure 161 parameters were identified as sensitive and the lower and upper limits of the searchspace were set to $p \pm 25\%$ where p is the initial parameter value.

5.1.2. The Reference Set

The reference set was build specifically to model the reversed AAC. The charges and molecular geometries in the reference set were calculated for a set of 1,2,3-triazoles and their respective dissociation products, the alkynes and azides, with various alcylic and arylic residues. They have been optimized on the RI-MP2/aug-cc-pVTZ level of theory and the partial charges were calculated using the CHELPG routine implemented in ORCA^[101,102] A total of 57 geometries were used in the optimization. The geometries used can be found in appendix C.

The energy landscape for the reversed AAC has been scanned in two dimensions on the B3LYP/6-31++G**level of theory. The scanned coordinates are the bond distances of the CN-bonds in the 1,2,3-triazole ring. The final version of the reference set contains 1531 single-point energies.

The reference set was then improved by adjusting the item weights in multiple preliminary optimization runs. The best set of weights was then used to obtain the best possible result for the *Mattsson* and the *Rahaman* force field as starting points.

5.1.3. Results

The optimization yielded two final parameter sets, one for the each initial force field chosen, i.e. *Mattsson's* and *Rahaman's* parameterizations. After the optimization the resulting force fields were compared directly with the reference data and applied for a small strained MD simulation.

Potential Energy Surface

The reference energies were comprised of a two-dimensional potential energy surface scan of the 1,2,3-triazole. All single-point energies have been compared with the REAXFF potential energies before and after the optimization of the force fields. The resulting deviation plots are shown in figure 5.1

In both cases the optimization smoothed out qualitative errors in the educt and the product region of the PES. When paying attention to the different color coding scalings of the energy difference ΔE it also stands out that the overall deviation of the energies has been reduced significantly in both cases.

The overall shape of the reparameterized *Rahaman* force field looks very promising. A minor quality issue with the surface is the overestimation of the reaction barrier height.

5. Parameterization of a 1,2,3-triazole force field

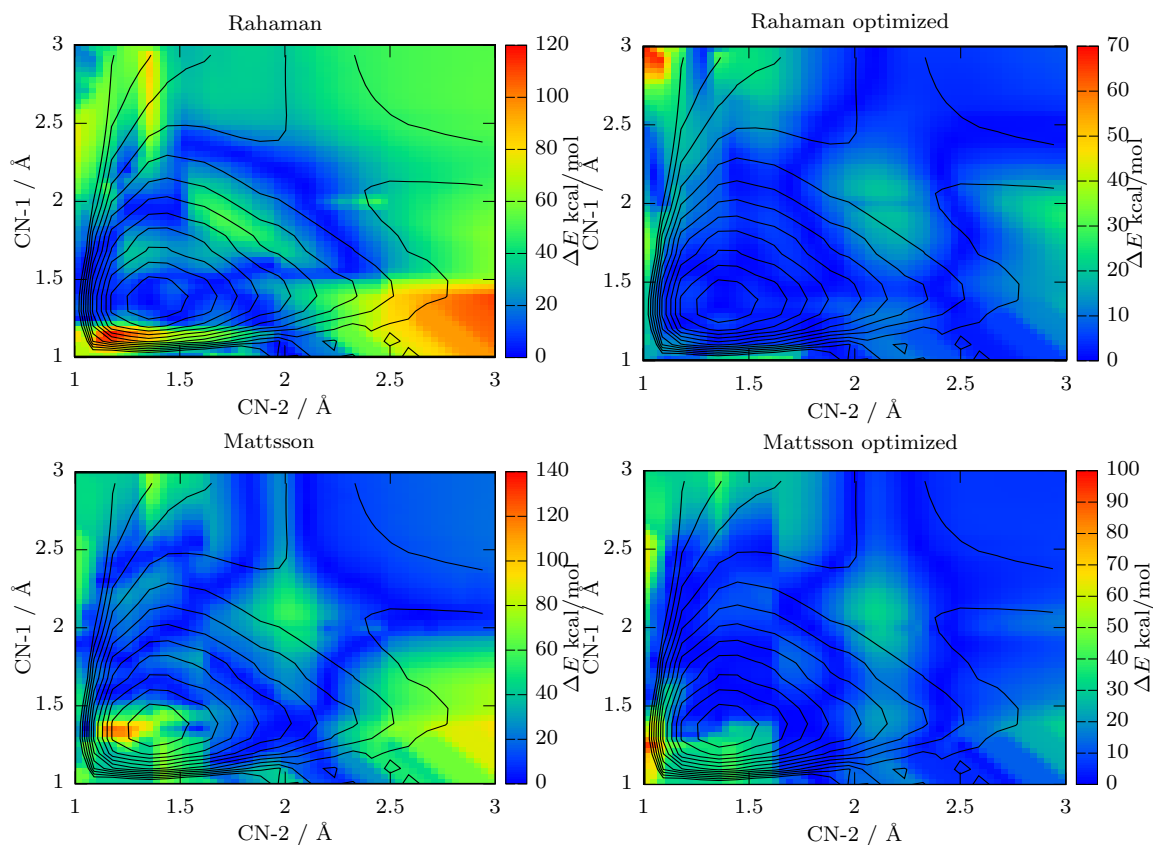


Figure 5.1.: Energy deviations of the force field energies and the reference DFT energies before and after reparameterization of the force fields by *Mattsson* and *Rahaman*. Only energy deviations are plotted, the underlying PES is indicated by the black isopotential lines. The 1,2,3-triazole educt state is located in the lower left corner of the PES while the dissociated products are found in the upper right corner. The figure is adapted from *Steffens F3 practical course thesis*^[4].

In case of the parameterization based on *Mattsson's* force field, larger deviations occur in the educt region after the optimization.

Molecular Geometries and Partial Charges

To assess the quality of the equilibrium structures with the optimized parameter sets, all reference geometries were optimized with both final solution candidates. Except for minor flaws in the NNN-angle of the azide product all geometries compared very well to the RI-MP2 reference.

The largest deviations occur in the molecular partial charges. As already mentioned in

previous chapters, there is no reliable way to determine partial charges. The CHELPG charges used already show unexpected behaviour. Therefore the QEq charges are deemed unproblematic as long as they are qualitatively correct and produce no computational artifacts.

5.2. Further Work

The parameter set obtained by *Steffen* was used to calculate the structural elongations of the 1,2,3-triazole system investigated two chapters before.

The simulations were done with the REAXFF implementation in the LAMMPS suite. The test structure was aligned for the anchoring atoms to line up with the x-axis of the coordinate system. The mechanophore in its closed and open form was then strained at low force loads of 0.5 nN and allowed to equilibrate at 300 K in the NVT-ensemble for 200 ps. After the equilibration a trajectory for both forms and the straining forces 1.11 nN, 1.21 nN and 2.05 nN each, which were the experimentally found breaking forces, was propagated for 100 ps. The distances between the anchoring atoms were averaged over the last 50 ps of each trajectory to obtain the structural lengths and consequentially the elongations upon rupture.

The resulting elongations are 8.7 Å at 1.11 nN, 8.8 Å at 1.21 nN and 10.9 Å at 2.05 nN. Again these results underestimate the elongation found experimentally by a significant margin. The error is even larger than that found during the DFT treatment. This underestimation is suspected to stem from an error along the potential energy curves which is discussed below.

As was already mentioned, the potential energy landscape after the optimization resembles the reference data closely for the transition region and the product and educt states. The problem with the parameterization, and most other REAXFF force fields for that matter, is that potential energy curves which are not part of the reference set are sometimes resembled poorly. Often times dissociation energies for single bonds are overestimated or underestimated significantly. This can lead to excessive gradients along the potential energy curves, the potential becomes harder (fig. 5.2). These harder potentials are more resilient against force-induced stretching, which in turn causes the underestimation in elongations that was pointed out before.

Part of the overestimation problem stems from the weighting coefficients in the error sum which were chosen for the single-point energies of the reference PES. They have been scaled with the potential energy to weight high-energy items less. Energies in dis-

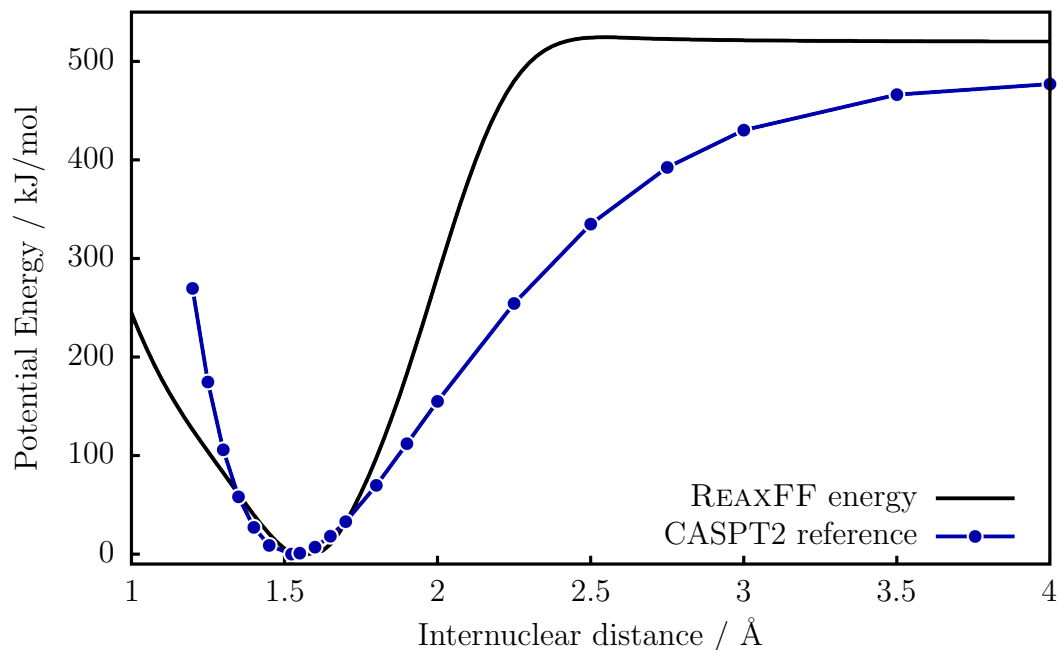


Figure 5.2.: Potential energies for the carbon-carbon single bond in ethane. The CASPT2 reference points (blue) are connected to guide the eye.

sociation regions and the transition state feel therefore less optimization pressure. This of course worked in favour of one of the quality criteria chosen in *Steffen's* work. Solution candidates were deemed promising during the manual curation when a certain amount of trajectories features no dissociation artifacts. These trajectories were propagated either in the educt or the product minima and therefore did not assess the force field quality at transition or dissociation regions.

A further set of MD simulations were done to investigate the dynamic behaviour of the 1,2,3-triazole mechanophore under an external straining force. The anchoring atoms were loaded with forces up to 6.5 nN upon which the system underwent spontaneous dissociation. In all reactive cases of a small set of 20 trajectories the PEG spacer chain disconnected from the central macrocycle. No reversed AAC or CN dissociation in the bridged region was observed.

After all the force field was planned to give mechanistic insight in the reversed AAC in solution and under the influence of external forces. The single-point energies were to be substituted with CASPT2 data to capture the static correlation in multireference region of the PES and get the relative energies quantitatively correct. But soon after it became clear that no reversed AAC is induced by the external force, the triazol param-

5. Parameterization of a 1,2,3-triazole force field

eterization project was abandoned. At the end of the day there is not much demand for parameterized empirical potentials for reactions that do not occur in experiments.

The bright side is that the parameterization by *Steffen* is another prime example for the ability of the REAXFF potential functions to model complex reaction energy surfaces in a very reasonable amount of time. With a little less than three months of work from start to finish the project demonstrates how well REAXFF can be applied to study covalent mechanochemical reactions.

Furthermore the obtained parameter set can be used as a starting point for the optimization of parameters for more promising 1,2,3-triazole derivatives. The already mentioned SPAAC is an example for a cycloreversion where considerably lower activation forces are expected. Another option is to change the 1,4-disubstituted 1,2,3-triazole with an 1,5-disubstituted one what increases the effective force F_{eff} acting on the reaction coordinate. Validation calculations of *Steffen's* parameter set need to reveal whether it is already capable of an accurate description of these reactions.

6. Summary and Outlook

First and foremost it was shown that REAXFF parameters can be globally optimized against high-quality multireference data by using evolutionary algorithms.

The parameterization we presented^[3] is, to the best of our knowledge, the first and only published REAXFF parameter set in the 15 year history of this force field that utilizes CASPT2 reference data. All other parameterizations rely on DFT reference data. For the description of CMC, DFT data is problematic due to its single-reference character. Mechanochemical reactivity crucially depends on the gradient vector field in the geometrical configuration space of a molecule, or more specifically regions of it where the gradient becomes large. The largest gradient along a dissociation curve, which is the bond breaking force, occurs at internuclear separations that are much larger than the equilibrium distance. In our publication we have shown that the multireference character of the wavefunction already becomes significant at these distances for the simple case of the disulfide dissociation curve^[3]. Single-reference methods like DFT are not reliable in these regions. In more complex cases like the 1,2,3-triazole treated by *Steffen*, where the concerted lengthening of two bonds happens simultaneously to the deformation of an aromatic system, the neglect of static correlation contributions may have dramatic effects on the curvature of the potential energy curves, and therefore on the dynamics of strained molecules in the potential.

Many published parameterizations, not only for reactive potentials but force fields in general, use small sized model systems as reference data. The general success of these methods indicate that large molecular systems can indeed be accurately modeled on the base of small molecule reference sets. Unfortunately due to the lack of experimental or high-quality ab-initio data for larger mechanophores it was not confirmed whether this is also true with CMC parameterizations for REAXFF. Nonetheless the disulfide parameterization is consistent with the models for CMC. This transferability from small molecule reference sets to large molecule MD simulations is a pleasant quality of parameterization based on multireference data, as MR methods are only feasible for small molecules.

6. Summary and Outlook

However, the global optimization of the REAXFF parameters revealed many hidden obstacles over the course of the project. In the early days of the project, progress was limited by the scarcity of reliable reference data as well as by the performance of the older EA implementation which was about one order of magnitude slower than the OGOLEM implementation^[1]. Due to the propagation behaviour of the EA on the objective function, which is fundamentally different from that of SOPPE or simulated annealing, the setup of reasonable search space boundaries was a challenging task. In the final phase of the reparameterization, overfitting became a serious issue. The guidelines and strategies applied to overcome all these challenges and to successfully optimize a force field for disulfide CMC are discussed at length in this thesis. Future parameterization projects, either for CMC or any other reaction, can build upon this knowledge to obtain results much faster.

The resulting force field was applied to model SMFS experiments on disulfides in proof-of-principle calculations. It was shown that the REAXFF formalism can be used to simulate the experiments and yields results that compare well to the kinetic models for mechanochemistry. The force field was successfully applied to calculate thousands of trajectories for a mechanophore in vacuo and thousands of trajectories for a mechanophore in a solvation box of toluene containing 2493 atoms on a small computation cluster¹. Furthermore a trajectory was propagated on a single core² for 0.5 μ s. With the excellent parallel scaling of the LAMMPS package, REAXFF is well suited to tackle single-molecular CMC on all relevant time- and system scales.

The timeframe for a reparameterization and its complexity are probably most critical when it comes to applicability for users who are not specialized in global optimization. Acknowledging this, techniques to fully or partially black-box the parameterization were developed. The practical course project by *Steffen* partially relied on a black-boxing approach, the reference set was assembled and weighted systematically and sensitive parameters indentified by the methods discussed in section 4.2.1. In less than three month it was possible to generate a stable force field from scratch. The less manually dependent approach chosen for the disulfide, which relies on trajectory data, was able to arrive at a parameter set in less than three weeks. The application of these black-box type methods is therefore comparably simple and fast. Depending on the available computational resources and the desired quality of the reference data, the parameterization for any given reaction can be done within a month.

¹About three weeks cluster walltime at 40 cores of AMD Opteron™ 2358 SE Quad-Core, 2.4 GHz and 48 cores of AMD Opteron™ 6172 12core, 2.1 GHz

²480h on a single workstation core Intel(R) Xeon(R) CPU E5-1620, 3.60GHz

6. Summary and Outlook

The real issue that remains with these black-box approaches is the quality and the quality assessment of the resulting parameter sets. While the approach by *Steffen* produced force fields that perform well in rMD applications, the disulfide force fields from the generic reference set approach failed even at small test calculations. Although these artifacts are obvious when a theoretician is looking for them, they are not easily identified by automatized routines. This goes against the idea of black-boxing the whole parameterization process.

Nonetheless the future perspective for such black-boxing approaches is positive. The reference set can be improved to better account for high-energy regions of the PES by using steered dynamics or systematic normal coordinate scans to generate the reference energies. The guidelines for setting up `params` files will improve the quality of resulting force fields. Current developments in automatic trajectory assessment may be used to define quality criteria for force fields that can be evaluated by a computer. Although the task is challenging to date, there is no practical reason for the black-boxing to be impossible in the near future.

In a broader sense, all of the above should be applicable not only to CMC but to all reactions in general. Furthermore, nothing about the approach restricts the application to REAXFF, the parameterization can be generalized to other reactive empirical potentials like COMB or the REBO family.

The second more application-oriented part of this thesis is concerned with mechanophores or molecules with predetermined breaking points in general. In the collaborative project A05 a macrocyclic 1,2,3-triazole with a bridged PBP was synthesized by organic chemists, investigated in SMFS experiments by physical chemists and computationally treated in the theoretical chemistry.

Unfortunately it was shown that the 1,2,3-triazole moiety is problematic in the configuration that was used in *Schütze's* SMFS experiments. The activation force for the reaction is too high to favour the reversed AAC in a strained setup over the homolytic fission of single bonds along the molecule.

Although these problems were present in the experiment, three successful force extension curves were obtained. The resulting structural elongation compared well enough to the computational models to classify the double rupture events as the expected two-step reaction where the first step is the rupture of the 1,2,3-triazole moiety and the second step the collapse of the remaining structure.

The activation force for the homolytic fission of the disulfide moiety was found to be significantly lower than for CC- or CO-bonds that are present in the anchoring poly-

6. Summary and Outlook

mer. A disulfide-based mechanophore is therefore expected to yield better results in SMFS experiments. The success rate for double ruptures in the force extension curves is expected to be higher due to the clear separation of the activation forces.

For future work the design of several different mechanophores is interesting. To empower ongoing work towards molecular logical structures or machines, a library of building blocks helps to construct them. These future mechanophoric building blocks should fulfill several demands. First, they should be able to be synthesized by known organic reactions. Second, and from the physicochemical experimental point of view much more important, they should feature low activation forces of less than 50 % of the weakest single bond in the system and a reaction coordinate parallel to the applied force. Only then the reaction of the mechanophore can be guaranteed.

Mechanophores already are used in experimental materials that show color changes upon mechanical stress which indicate material failure. The bridging of the mechanically active unit may add a reversibility aspect to these application. These may then be used in regenerative polymeres or sensor applications.

Bibliography

- [1] M. Dittner, J. Müller, M. Aktulga, and B. Hartke. *J. Comput. Chem.*, 36:1550, 2015.
- [2] D. Schütze, K. Holz, J. Müller, M. K. Beyer, U. Lüning, and B. Hartke. *Angew. Chem. Int. Ed.*, 54:2556, 2015.
- [3] J. Müller and B. Hartke. *J. Chem. Theo. Comput.*, 12:3913, 2016.
- [4] J. Steffen. *Globale Parameteroptimierung des Reaktiven Kraftfeldes ReaxFF für die Cycloreversion von Triazolsystemen, F3 Thesis*. CAU Kiel, 2015.
- [5] G. Binnig, C. F. Quate, and Ch. Gerber. *Phys. Rev. Lett.*, 56:930, 1986.
- [6] M. Rief, F. Oesterhelt, B. Heymann, and H. E. Gaub. *Science*, 275:1295, 1997.
- [7] M. Granbois, M. K. Beyer, M. Rief, H. Clausen-Schaumann, and H. E. Gaub. *Science*, 283:1727, 1999.
- [8] J. N. Brantley, K. M. Wiggins, and C. W. Bielawski. *Science*, 333:1606, 2011.
- [9] M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour. *Science*, 278:252, 1997.
- [10] P. A. Wiita, S. R. K. Ainavarapu, H. H. Huang, and J. M. Fernandez. *Proc. Natl. Acad. Sci.*, 103:7222, 2006.
- [11] M. K. Beyer, U. Lüning, and B. Hartke. http://www.sfb677.uni-kiel.de/pages_en/projekte_a05_index.html, 2011. Online; accessed 28-July-2016.
- [12] A. C. T van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard III. *J. Phys. Chem. A*, 105:9396, 2001.
- [13] J. Ribas-Ariño and D. Marx. *Chem. Rev.*, 112:5412, 2012.

Bibliography

- [14] L. Takacs. *Chem. Soc. Rev.*, 42:7649, 2013.
- [15] W. Ostwald. *Lehrbuch der Allgemeinen Chemie vol. 2, part 1*. Verlag von Wilhelm Engelmann, Leipzig, 1903.
- [16] M. Sopicka-Lizer. *High-Energy Ball Milling: Mechanochemical Processing of Nanopowders*. Woodhead Publishing, 2010.
- [17] D. Chen, S. K. Sharma, and A. Mudhoo. *Handbook on Applications of Ultrasound: Sonochemistry for Sustainability*. CRC Press, 2011.
- [18] A. L. Black, J. M. Lenhardt, and S. L. Craig. *J. Mat. Chem.*, 21:1655, 2011.
- [19] H. Eyring, J. Walter, and G. E. Kimball. *Quantum Chemistry*. John Wiley & Sons, 1944.
- [20] S. N. Zhurkov. *Int. J. Frac.*, 1:311, 1965.
- [21] G. I. Bell. *Science*, 200:618, 1978.
- [22] H. S. Smalø, V. V. Rybkin, W. Klopper, T. Helgaker, and E. Uggerud. *J. Phys. Chem. A*, 118:7683, 2014.
- [23] S. S. M. Konda, J. N. Brantley, C. W. Bielawski, and D. E. Makarov. *J. Chem. Phys.*, 135:164103, 2011.
- [24] M. K. Beyer. *J. Chem. Phys.*, 112:7303, 2000.
- [25] E. Evans and K. Ritchie. *Biophys. J.*, 72:1541, 1997.
- [26] L. B. Freund. *Proc. Nat. Acad. Sci.*, 106:8819, 2009.
- [27] S. M. Avdoshenko and D. E. Makarov. *J. Phys. Chem. B*, 120:1537, 2016.
- [28] E. Evans. *Faraday Discuss.*, 111:1, 1998.
- [29] J. Ribas-Ariño, M. Shiga, and D. Marx. *Angew. Chem.*, 121:4254, 2009.
- [30] TURBOMOLE V6.2 2010, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from <http://www.turbomole.com>.

Bibliography

- [31] S. J. Plimpton. <http://lammps.sandia.gov/>, 1995. Online; accessed 4-March-2017.
- [32] L. Verlet. *Phys. Rev.*, 159:98, 1967.
- [33] L. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande, and T. J. Martínez. *Nat. Chem.*, 6:1044, 2014.
- [34] J. R. Maple, Hwang M. J., T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewig, and A. T. Hagler. *J. Comput. Chem.*, 15:162, 1994.
- [35] F. Jensen. *Introduction to Computational Chemistry*. Wiley, 2007.
- [36] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. *J. Am. Chem. Soc.*, 117:11225, 1996.
- [37] J. Ponder. <https://dasher.wustl.edu/tinker/>, 1991. Online; accessed 4-March-2017.
- [38] S. A. Adcock and J. A. McCammon. *Chem. Rev.*, 106:1589, 2006.
- [39] A. Warshel and R. M. Weiss. *J. Am. Chem. Soc.*, 102:6218, 1980.
- [40] P. L. A. Popelier. *Int. J. Quant. Chem.*, 115:2015, 2015.
- [41] J. T. Su and W. A. Goddard III. *Phys. Rev. Lett.*, 99:185003, 2007.
- [42] T. Liang, Y. K. Shin, Y. Cheng, D. E. Yilmaz, K. G. Vishnu, O. Varners, C. Zou, S. R. Phillpot, S. B. Sinnott, and A. C. T. van Duin. *Ann. Rev. Mater. Res.*, 43:109, 2013.
- [43] M. W. Finnis and J. E. Sinclair. *Philosophical Magazine A*, 50:45, 1984.
- [44] S. D. Murray and M. I. Baskes. *Phys. Rev. B*, 29:6443, 1984.
- [45] M. I. Baskes. *Phys. Rev. Lett.*, 59:2666, 1987.
- [46] G. C. Abell. *Phys. Rev. B*, 31:6184, 1985.
- [47] J. Tersoff. *Phys. Rev. B*, 37:6991, 1988.
- [48] D. W. Brenner. *Phys. Rev. B*, 42:9485, 1990.

Bibliography

- [49] D. W. Brenner, O. A. Shenderova, J. A. Harrison, S. J. Stuart, and S. B. Sinnott. *J. Phys. Condens. Matter*, 14:783, 2002.
- [50] S. J. Stuart, A. B. Tutein, and J. A. Harrison. *J. Chem. Phys.*, 112:6472, 2000.
- [51] S. Grimme. *J. Chem. Theory Comput.*, 10:4497, 2014.
- [52] B. Hartke and S. Grimme. *Phys. Chem. Chem. Phys.*, 17:16715, 2015.
- [53] J. Steffen and B. Hartke. *J. Chem. Phys.*, 147:161701, 2017.
- [54] W. J. Mortier, S. W. Ghosh, and S. Shankar. *J. Am. Chem. Soc.*, 108:4315, 1986.
- [55] A. K. Rappe and W. A. Goddard III. *J. Phys. Chem.*, 95:3358, 1991.
- [56] A. C. T. van Duin, A. Strachan, S. Stewman, Q. Zhang, X. Xu, and W. A. Goddard III. *J. Phys. Chem. A*, 107:3803, 2003.
- [57] K. D. Nielson, A. C. T. van Duin, J. Oxgaard, W. Deng, and W. A. Goddard III. *J. Chem Phys. A*, 109:493, 2005.
- [58] K. Chenoweth, A. C. T. van Duin, and W. A. Goddard III. *J. Phys. Chem A*, 112:1040, 2008.
- [59] L. Liu, Y. Lu, S. V. Zybin, H. Sun, and W. A. Goddard III. *J. Phys. Chem.*, 115:11016, 2011.
- [60] H. M. Aktulga, J. C. Fogarty, S. A. Pandit, and A. Y. Grama. *Parallel Comp.*, 38:245, 2012.
- [61] L. Pauling. *J. Am. Chem. Soc.*, 69:542, 1947.
- [62] K. Joshi, A. C. T. van Duin, and T. Jacob. *J. Mater. Chem.*, 20:10431, 2010.
- [63] T. Weise. *Global Optimization Algorithms - Theory and Application* -. selfpublished, 2011.
- [64] E. Iype, M. Hütter, A. P. J. Jansen, S. V. Nedea, and C. C. M. Rindt. *J. Comput. Chem.*, 34:1134, 2013.
- [65] J. P. Larentzos, B. M. Rice, E. F. C. Byrd, N. S. Weingarten, and J. V. Lill. *J. Chem. Theory Comput.*, 11:381, 2015.

Bibliography

- [66] D. H. Wolpert and W. G. Macready. *Technical Report SFI-TR-95-02-010, Santa Fe Institute*, 10:1, 1995.
- [67] J. M. Dieterich and B. Hartke. <https://www.ogolem.org/>, 2010. Online; accessed 11-March-2017.
- [68] J. M. Dieterich and B. Hartke. *Mol. Phys.*, 108:279, 2010.
- [69] T. Weise, R. Chiong, and K. Tang. *J. Comput. Sci.*, 27:907, 2011.
- [70] D. E. Goldberg. *Genetic Algorithms in Search Optimization & Machine Learning*. Addison Wesley, 1989.
- [71] B. Hartke. *J. Comput. Chem.*, 99:161752, 1999.
- [72] M. Dittner. *Neue Implementation globaler Parameteroptimierung eines reaktiven Kraftfeldes, M.Sc Thesis*. CAU Kiel, 2014.
- [73] B. Bandow and B. Hartke. *J. Phys. Chem. A*, 110:5809, 2006.
- [74] J. M. Dieterich and B. Hartke. *J. Chem. Theory Comput.*, 12:5226, 2016.
- [75] E. Schrödinger. *Phys. Rev.*, 28:1049, 1926.
- [76] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry*. Dover Books on Chemistry, 1996.
- [77] W. Kutzelnigg. *Einführung in die Theoretische Chemie*. Wiley-VCH, 2001.
- [78] T. Helgaker, J. Olsen, and P. Jorgensen. *Molecular Electronic-Structure Theory*. Wiley, 2013.
- [79] B. O. Roos, R. Lindh, P. Å. Malmqvist, V. Veryazov, and P. O. Widmark. *Multi-configurational Quantum Chemistry*. Wiley, 2016.
- [80] R. Huisgen. *Proc. Chem. Soc.*, page 357, 1961.
- [81] M. F. Pill, K. Holz, N. Preußke, F. Berger, H. Clausen-Schaumann, U. Lüning, and M. K. Beyer. *Chem. Eur. J.*, 22:12034, 2016.
- [82] A. Schäfer, C. Huber, and R. Ahlrichs. *J. Chem. Phys.*, 100:5829, 1994.

Bibliography

- [83] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs. *Chem. Phys. Lett.*, 240:283, 1995.
- [84] K. Eichkorn, F. Weigend, O. Treutler, and R. Ahlrichs. *Theor. Chem. Acc.*, 97:119, 1997.
- [85] F. Weigend. *Phys. Chem. Chem. Phys.*, 8:1057, 2005.
- [86] M. von Arnim and R. Ahlrichs. *J. Comput. Chem.*, 19:1746, 1998.
- [87] M. von Arnim and R. Ahlrichs. *J. Chem. Phys.*, 111:9183, 1999.
- [88] F. Weigend. *Phys. Chem. Chem. Phys.*, 4:4285, 2002.
- [89] R. Ahlrichs. *Phys. Chem. Chem. Phys.*, 6:5119, 2004.
- [90] J. Ribas-Ariño, M. Shiga, and D. Marx. *J. Am. Chem. Soc.*, 132:10609, 2010.
- [91] M. J. Jacobs, G. Schneider, and K. G. Blank. *Angew. Chem.*, 128:2950, 2016.
- [92] J. M. Baskin, J. A. Prescher, S. T. Laughlin, N. J. Agard, P. V. Chang, I. A. Miller, A. Lo, J. A. Codelli, and C. R. Bertozzi. *Proc. Natl. Acad. Sci.*, 104:16793, 2007.
- [93] C. R. Becer, R. Hoogenboom, and U. S. Schubert. *Angew. Chem.*, 121:4998, 2009.
- [94] C. Peng and H. B. Schlegel. *Israel Journal of Chemistry*, 33:449, 1993.
- [95] 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, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, , D. J. Fox, Gaussian Inc., and Wallingford CT 2016., 2009.

Bibliography

- [96] M. T. Ong, J. Leiding, H. Tao, A. M. Virshup, and T. J. Martinez. *J. Am. Chem. Soc.*, 131:6377, 2009.
- [97] M. F. Pill, S. W. Schmidt, M. K. Beyer, H. Clausen-Schaumann, and A. Kersch. *J. Chem. Phys.*, 140:044321, 2014.
- [98] C Jia, Migliore A., N. Xin, S. Huang, J. Wang, Q. Yang, S. Wang, H. Chen, D. Wang, Feng. B., Z. Liu, G. Zhang, D. Qu, H. Tian, M. A. Ratner, H. Q. Xu, A. Nitzan, and X. Gui. *Science*, 352:1443, 2016.
- [99] H. R. Larsson, A. C. T. van Duin, and B. Hartke. *J. Comput. Chem.*, 34:2178, 2013.
- [100] H. R. Larsson and B. Hartke. *CMMS*, 13:120, 2013.
- [101] F. Neese. *WIREs Comput. Mol. Sci.*, 2:73, 2012.
- [102] F. Neese. *J. Comput. Chem.*, 24:1740, 2003.
- [103] Daylight Chemical Information Systems Inc. <http://daylight.com/smiles/>, 2008. Online; accessed 20-April-2017.
- [104] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, and G. R. Vandermeersch, T. Hutchison. http://openbabel.org/wiki/Main_Page, 2005. Online; accessed 20-April-2017.
- [105] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, and G. R. Vandermeersch, T. Hutchison. *J. of Cheminformatics*, 3:33, 2011.
- [106] A. C. T. van Duin. <https://www.scm.com/wp-content/uploads/ReaxFF-users-manual-2002.pdf>, 2002. Online; accessed 5-April-2017.
- [107] T. R. Mattsson, J. M. D. Lane, K. R. Cochrane, M. P. Desjarlais, A. P. Thompson, F. Pierce, and G. S. Grest. *Phys. Rev. B*, 81:054103, 2010.
- [108] O. Rahaman, A. C. T. van Duin, W. A. Goddard III., and D. J. Doren. *J. Phys. Chem. B*, 115:249, 2012.
- [109] J. E. Mueller, A. C. T. van Duin, and W. A. Goddard III. *J. Phys. Chem.*, 114:4939, 2010.

Appendices

A. ReaxFF Optimization Input Files

The following section contains examples for every input file used in the optimization of a REAXFF parameter set with OGOLEM. All examples are chosen from the input set used for the disulfide parameter optimization published in 2016^[3].

A.1. The ffield File

The example force field is the final parameter set obtained with the global optimization for the disulfide system.

DATE: 2016-02-19 CITATION: Mueller, J. and Hartke, B., JCTC (2016), 12, 3913

```
39      ! Number of general parameters
    50.0000 !Overcoordination parameter
    9.5469 !Overcoordination parameter
127.8302 !Valency angle conjugation parameter
    3.0000 !Triple bond stabilisation parameter
    6.5000 !Triple bond stabilisation parameter
    0.0000 !C2-correction
    1.0496 !Undercoordination parameter
    9.0000 !Triple bond stabilisation parameter
11.5054 !Undercoordination parameter
13.4059 !Undercoordination parameter
    0.0000 !Triple bond stabilization energy
    0.0000 !Lower Taper-radius
10.0000 !Upper Taper-radius
    2.8793 !Not used
33.8667 !Valency undercoordination
    7.0994 !Valency angle/lone pair parameter
    1.0563 !Valency angle
    2.0384 !Valency angle parameter
    6.1431 !Not used
    6.9290 !Double bond/angle parameter
    0.3989 !Double bond/angle parameter: overcoord
    3.9954 !Double bond/angle parameter: overcoord
```

A. REAXFF Optimization Input Files

```

-2.4837 !Not used
 5.7796 !Torsion/BO parameter
10.0000 !Torsion overcoordination
 1.9487 !Torsion overcoordination
-1.2327 !Conjugation 0 (not used)
 2.1645 !Conjugation
 1.5591 !vdWaals shielding
 0.1000 !Cutoff for bond order (*100)
 2.0038 !Valency angle conjugation parameter
 0.6121 !Overcoordination parameter
 1.2172 !Overcoordination parameter
 1.8512 !Valency/lone pair parameter
 0.5000 !Not used
20.0000 !Not used
 5.0000 !Molecular energy (not used)
 0.0000 !Molecular energy (not used)
 3.6942 !Valency angle conjugation parameter
4   ! Nr of atoms; cov.r; valency;a.m;Rvdw;Evdw;gammaEEM;cov.r2;#
      alfa;gammavdW;valency;Eunder;Eover;chiEEM;etaEEM;n.u.
      cov r3;Elp;Heat inc.;n.u.;n.u.;n.u.;n.u.
      ov/un;val1;n.u.;val3,vval4
C   1.3763  4.0000 12.0000  1.8857  0.1818  0.8712  1.2596  4.0000
      9.5928  1.6819  4.0000 42.7976 79.5548  5.7254  6.9235  0.0000
      1.2065  0.0000 -0.8579  8.7956 19.0071 21.6867  0.8563  0.0000
      -7.7789  3.2369  1.0564  4.0000  2.8623  0.0000  0.0000  0.0000
H   0.6646  1.0000  1.0080  1.6030  0.0600  0.7625 -0.1000  1.0000
      9.3951  4.5386  1.0000  0.0000 121.1250  3.8196  9.8832  1.0000
      -0.1000  0.0000 -0.1339  2.5732  2.6456  3.1680  1.0698  0.0000
      -12.9330  3.0626  1.0338  1.0000  2.8793  0.0000  0.0000  0.0000
O   1.2699  2.0000 15.9990  1.9741  0.0880  1.0804  1.0624  6.0000
      10.2186  7.7719  4.0000 27.3264 116.0768  8.5000  7.8386  2.0000
      0.9446  8.6170 -1.2371 17.0845  3.7082  0.5350  0.9745  0.0000
      -3.1456  2.6656  1.0493  4.0000  2.9225  0.0000  0.0000  0.0000
S   1.6951  2.0000 32.0600  1.9019  0.6725  1.0336 -0.1000  6.0000
      9.6692  4.9160  4.0000 55.8316 112.1416  6.5000  8.2545  2.0000
      -0.1000  9.7177 -2.3700 16.9855 12.7440  3.0488  0.9745  0.0000
      -9.0708  3.7542  1.0338  4.0000  2.8956  0.0000  0.0000  0.0000
10  ! Nr of bonds; Edis1;LPpen;n.u.;pbe1;pbo5;13corr;pbo6
      pbe2;pbo3;pbo4;n.u.;pbo1;pbo2;ovcorr
1  1 145.4070 103.0681 73.7841 0.2176 -0.7816  1.0000 28.4167  0.3217
      0.1111 -0.1940  8.6733  1.0000 -0.0994  5.9724  1.0000  0.0000
1  2 167.1752  0.0000  0.0000 -0.4421  0.0000  1.0000  6.0000  0.5969

```

A. REAXFF Optimization Input Files

```

17.4194 1.0000 0.0000 1.0000 -0.0099 8.5445 0.0000 0.0000
1 3 171.0470 67.2480 130.3792 0.3600 -0.1696 1.0000 12.0338 0.3796
0.3647 -0.2660 7.4396 1.0000 -0.1661 5.0637 0.0000 0.0000
1 4 123.5848 0.0000 0.0000 0.3109 0.0000 1.0000 6.0000 0.5376
8.9045 1.0000 0.0000 1.0000 -0.1597 4.5508 1.0000 0.0000
2 2 188.1606 0.0000 0.0000 -0.3140 0.0000 1.0000 6.0000 0.6816
8.6247 1.0000 0.0000 1.0000 -0.0183 5.7082 0.0000 0.0000
2 3 216.6018 0.0000 0.0000 -0.4201 0.0000 1.0000 6.0000 0.9143
4.7737 1.0000 0.0000 1.0000 -0.0591 5.9451 0.0000 0.0000
2 4 153.2367 0.0000 0.0000 -0.3890 0.0000 1.0000 6.0000 0.6485
3.4079 1.0000 0.0000 1.0000 -0.1395 5.6749 1.0000 0.0000
3 3 90.2465 160.9645 40.0000 0.9950 -0.2435 1.0000 28.1614 0.9704
0.8145 -0.1850 7.5281 1.0000 -0.1283 6.2396 1.0000 0.0000
3 4 0.0000 0.0000 0.0000 0.5563 0.0000 1.0000 6.0000 0.6000
0.4259 -0.4577 12.7569 1.0000 -0.1100 7.1145 1.0000 0.0000
4 4 117.1855 0.0000 0.0000 0.5590 0.0000 1.0000 6.0000 0.5192
3.9845 1.0000 0.0000 1.0000 -0.2515 4.2968 1.0000 0.0000
6 ! Nr of off-diagonal terms; Ediss;Ro;gamma;rsigma;rpi;rpi2
1 2 0.0455 1.7218 10.4236 1.0379 -1.0000 -1.0000
1 3 0.1186 1.9820 9.5927 1.2936 1.1203 1.0805
1 4 0.3463 1.8985 9.6518 1.5209 -1.0000 -1.0000
2 3 0.0469 1.9185 10.3707 0.9406 -1.0000 -1.0000
2 4 0.3157 1.3772 9.9744 1.3300 -1.0000 -1.0000
3 4 0.1359 2.0203 10.1000 1.6050 1.3050 -1.0000
31 ! Nr of angles;at1;at2;at3;Thetao,o;ka;kb;pv1;pv2
1 1 1 70.0265 13.6338 2.1884 0.0000 0.1676 26.3587 1.0400
1 1 2 69.7786 10.3544 8.4326 0.0000 0.1153 0.0000 1.0400
1 1 3 72.9588 16.7105 3.5244 0.0000 1.1127 0.0000 1.1880
1 1 4 72.6832 38.5451 4.1051 0.1463 1.1777 0.0000 2.0130
1 2 1 0.0000 3.4110 7.7350 0.0000 0.0000 0.0000 1.0400
1 3 1 79.1091 45.0000 0.7067 0.0000 0.6142 0.0000 1.0783
1 3 2 78.1533 44.7226 1.3136 0.0000 0.1218 0.0000 1.0500
1 3 3 83.7151 42.6867 0.9699 0.0000 0.6142 0.0000 1.0783
1 3 4 85.3644 36.9951 2.0903 0.1463 0.0559 0.0000 1.0400
1 4 1 80.9601 35.9462 0.9508 0.1463 2.1025 0.0000 1.3765
1 4 2 86.1791 36.9951 2.0903 0.0000 0.0000 0.0000 1.0400
1 4 3 85.3644 36.9951 2.0903 0.1463 0.0559 0.0000 1.0400
1 4 4 79.4158 21.9773 1.9571 0.1463 2.1939 0.0000 1.5061
2 1 2 74.6020 11.8629 2.9294 0.0000 0.1367 0.0000 1.0400
2 1 3 66.6150 13.6403 3.8212 0.0000 0.0755 0.0000 1.0500
2 1 4 74.9397 25.0560 1.8787 0.0000 0.0000 0.0000 1.0400
2 3 2 79.2954 26.3838 2.2044 0.0000 0.1218 0.0000 1.0500

```


A. REAXFF Optimization Input Files

```
2 3 3 84.1057 9.6413 7.5000 0.0000 0.1218 0.0000 1.0500
2 3 4 84.1057 9.6413 7.5000 0.0000 0.1218 0.0000 1.0500
2 4 2 90.4790 41.4332 0.5118 0.0000 1.8940 0.0000 1.0400
2 4 3 84.3331 36.9951 2.0903 0.0000 0.0000 0.0000 1.0400
2 4 4 73.3934 13.8498 1.7976 0.0000 1.5770 0.0000 1.0400
3 1 3 80.0708 45.0000 2.1487 0.0000 1.1127 -35.0000 1.1880
3 1 4 80.0708 45.0000 2.1487 0.0000 1.1127 -35.0000 1.1880
3 3 3 80.0108 38.3716 1.1572 -38.4200 0.6142 0.0000 1.0783
3 3 4 80.0108 38.3716 1.1572 -38.4200 0.6142 0.0000 1.0783
3 4 3 80.0108 38.3716 1.1572 -38.4200 0.6142 0.0000 1.0783
3 4 4 80.0108 38.3716 1.1572 -38.4200 0.6142 0.0000 1.0783
4 1 4 71.4075 21.3515 1.5584 0.0000 2.0477 -35.0000 2.2611
4 3 4 80.0108 38.3716 1.1572 -38.4200 0.6142 0.0000 1.0783
4 4 4 80.0108 38.3716 1.1572 -38.4200 0.6142 0.0000 1.0783
15 ! Nr of torsions;at1;at2;at3;at4;;V1;V2;V3;V2(B0);vconj;n.u;n
1 1 1 1 0.0000 23.2168 0.1811 -4.6220 -1.9387 0.0000 0.0000
1 1 1 2 0.0000 45.7984 0.3590 -5.7106 -2.9459 0.0000 0.0000
2 1 1 2 0.0000 44.6445 0.3486 -5.1725 -0.8717 0.0000 0.0000
1 1 4 4 3.3423 30.3435 0.0365 -2.7171 0.0000 0.0000 0.0000
1 4 4 1 6.2190 -15.0361 -0.8774 -1.5724 0.0000 0.0000 0.0000
0 1 1 0 0.0000 23.2168 0.1811 -4.6220 -1.9387 0.0000 0.0000
0 1 2 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
0 1 3 0 5.0520 16.7344 0.5590 -3.0181 -2.0000 0.0000 0.0000
0 1 4 0 3.3423 30.3435 0.0365 -2.7171 0.0000 0.0000 0.0000
0 2 2 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
0 2 3 0 0.0000 0.1000 0.0200 -2.5415 0.0000 0.0000 0.0000
0 2 4 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
0 3 3 0 0.0115 68.9706 0.8253 -28.4693 0.0000 0.0000 0.0000
0 3 4 0 0.0115 68.9706 0.8253 -28.4693 0.0000 0.0000 0.0000
0 4 4 0 0.7048 -20.6682 1.2272 -0.9578 0.0000 0.0000 0.0000
4 ! Nr of hydrogen bonds;at1;at2;at3;Rhb;Dehb;vhb1
3 2 3 2.0431 -6.6813 3.5000 1.7295
3 2 4 2.6644 -3.9547 3.5000 1.7295
4 2 3 2.1126 -4.5790 3.5000 1.7295
4 2 4 1.9461 -4.0000 3.5000 1.7295
```

A.2. The geo File

The example geo file contains three different dimethyl disulfide geometries in the BIOGRF format. The DESCRP flag is followed by a unique identifier to which the reference items in the `trainset.in` refer. The runtime specified by the RUTYPE calls either for a local

A. REAXFF Optimization Input Files

optimization (NORMAL RUN) during the objective evaluation or just an energy calculation (SINGLE POINT).

```

BIOGRF 200
DESCRP dmds
RUTYPE NORMAL RUN
FORMAT ATOM (a6,1x,i5,1x,a5,1x,a3,1x,a1,1x,a5,3f10.5,1x,a5,i3,i2,1x,f8.5)
HETATM  1 C1  RES A  444  1.47738 -0.02026  0.05289 C_3  4 0 -0.00782
HETATM  2 S2  RES A  444  2.73288 -0.96190 -0.85500 S_3  1 0 -0.09678
HETATM  3 S3  RES A  444  3.41658 -2.24371  0.58516 S_3  1 0 -0.09678
HETATM  4 C4  RES A  444  4.67118 -1.23232  1.41599 C_3  4 0 -0.00782
HETATM  5 H5  RES A  444  1.03289  0.67032 -0.66250 H_   1 0  0.03487
HETATM  6 H6  RES A  444  1.92820  0.54685  0.86270 H_   1 0  0.03487
HETATM  7 H7  RES A  444  0.71347 -0.68818  0.43933 H_   1 0  0.03487
HETATM  8 H8  RES A  444  5.11564 -1.86276  2.18492 H_   1 0  0.03487
HETATM  9 H9  RES A  444  4.21970 -0.36255  1.88538 H_   1 0  0.03487
HETATM 10 H10 RES A  444  5.43526 -0.92536  0.70819 H_   1 0  0.03487
FORMAT CONECT
END

```

```

BIOGRF 200
DESCRP dmds-SS2.042
RUTYPE SINGLE POINT
FORMAT ATOM (a6,1x,i5,1x,a5,1x,a3,1x,a1,1x,a5,3f10.5,1x,a5,i3,i2,1x,f8.5)
HETATM  1 C1  RES A  444  0.00000  0.00000  0.00000 C    1 0  0.00000
HETATM  2 C2  RES A  444  2.15854  1.76834  2.39608 C    1 0  0.00000
HETATM  3 S3  RES A  444  1.81307  0.00000  0.00000 S    1 0  0.00000
HETATM  4 S4  RES A  444  2.23109  0.00000  2.00245 S    1 0  0.00000
HETATM  5 H5  RES A  444 -0.30823 -0.03675 -1.04398 H    1 0  0.00000
HETATM  6 H6  RES A  444 -0.38786  0.90765  0.45429 H    1 0  0.00000
HETATM  7 H7  RES A  444 -0.37560 -0.87427  0.52307 H    1 0  0.00000
HETATM  8 H8  RES A  444  2.40871  1.85954  3.45218 H    1 0  0.00000
HETATM  9 H9  RES A  444  1.15914  2.16167  2.23138 H    1 0  0.00000
HETATM 10 H10 RES A  444  2.88266  2.31577  1.80025 H    1 0  0.00000
FORMAT CONECT (a6,12i6)
END

```

```

BIOGRF 200
DESCRP dmds-SS2.5
RUTYPE SINGLE POINT
FORMAT ATOM (a6,1x,i5,1x,a5,1x,a3,1x,a1,1x,a5,3f10.5,1x,a5,i3,i2,1x,f8.5)
HETATM  1 C1  RES A  444  0.00000  0.00000  0.00000 C    1 0  0.00000
HETATM  2 C2  RES A  444  2.25140  1.76834  2.84087 C    1 0  0.00000

```

A. REAXFF Optimization Input Files

```
HETATM 3 S3 RES A 444 1.81307 0.00000 0.00000 S 1 0 0.00000
HETATM 4 S4 RES A 444 2.32394 0.00000 2.44725 S 1 0 0.00000
HETATM 5 H5 RES A 444 -0.30823 -0.03675 -1.04398 H 1 0 0.00000
HETATM 6 H6 RES A 444 -0.38786 0.90765 0.45429 H 1 0 0.00000
HETATM 7 H7 RES A 444 -0.37560 -0.87427 0.52307 H 1 0 0.00000
HETATM 8 H8 RES A 444 2.50157 1.85954 3.89698 H 1 0 0.00000
HETATM 9 H9 RES A 444 1.25200 2.16167 2.67617 H 1 0 0.00000
HETATM 10 H10 RES A 444 2.97551 2.31577 2.24505 H 1 0 0.00000
FORMAT CONECT (a6,12i6)
END
```

A.3. The `trainset.in` File

The `trainset.in` file contains the actual reference data. For better readability of the reference set it is possible to include gradient information as a path to external gradient files.

The first block of the reference set refers to the geometry of the `dmds` molecule. The MP2 reference structure is provided as full z -matrix. Bonds, angles and dihedrals are referred to by the atoms they entail. The column after the geometry identifier is the weight of the property for the objective function. The units are Ångstrom and degree, respectively.

The second block contains gradient information. In this case the gradient is given by the molecule identifier, the weight of the gradient and the path to the file which contains the actual gradient vector for the molecule.

The energy section contains relative energies between structures in the `geo` file. The first column is the weight for the objective function followed by an expression that defines which relative energy is to be calculated. The last column carries the reference energy in kcal/mol.

```
GEOMETRY
dmds 0.02 2 1 1.813
dmds 0.01 3 2 2.046
dmds 0.02 4 3 1.813
dmds 0.05 5 1 1.089
dmds 0.05 6 1 1.087
dmds 0.05 7 1 1.086
dmds 0.05 8 4 1.089
dmds 0.05 9 4 1.087
dmds 0.05 10 4 1.086
```

A. REAXFF Optimization Input Files

```
dmds 3.00 1 2 3 101.791
dmds 3.00 2 3 4 101.791
dmds 3.00 2 1 5 106.439
dmds 3.00 2 1 6 110.914
dmds 3.00 2 1 7 110.238
dmds 3.00 3 4 8 106.439
dmds 3.00 3 4 9 110.913
dmds 3.00 3 4 10 110.238
dmds 3.00 1 2 3 4 -85.105
dmds 3.00 3 2 1 5 -177.984
dmds 3.00 3 2 1 6 63.411
dmds 3.00 3 2 1 7 -59.108
dmds 3.00 2 3 4 8 -177.984
dmds 3.00 2 3 4 9 63.411
dmds 3.00 2 3 4 10 -59.108
ENDGEOMETRY
GRADIENT
dmds-SS2.042 0.01 grads/dmds-SS2.042.gradient
dmds-SS2.5 0.01 grads/dmds-SS2.5.gradient
ENDGRADIENT
ENERGY
1.00 + dmds-SS2.5/1 - dmds-SS2.042/1 20.63543
ENDENERGY
```

The file `dmds-SS2.5.gradient` that contains the gradient of the structure `dmds-SS2.5` is constructed similar to an `xyz`-file. The gradient has the unit Hartree/Bohr. The atoms are addressed by their line number in the `geo` file. Note that not all atoms need to be specified, if the user is only interested in a small subset of the gradient vector it is possible to use only that.

10

```
1 -0.000866 -0.000439 -0.007328
2 0.002003 -0.000910 0.007181
3 -0.011944 -0.002635 -0.046176
4 0.012041 0.001972 0.045980
5 -0.001390 0.000299 0.001053
6 0.000348 -0.000102 0.000459
7 0.000765 0.000140 0.000685
8 -0.000631 0.001676 -0.000849
9 -0.000022 0.000307 -0.000313
10 -0.000305 -0.000307 -0.000692
```

A.4. The params File

The `params` file used for the disulfide optimization contain 87 parameters. They are identified by three integer numbers as discussed in section 4.4.1. This address is followed by a number which is not relevant for the EA optimization and only kept for compatibility with SOPPE. The last two columns hold the minimal and the maximal value for the parameter during the optimization which directly determines the size of the search space, for each parameter.

```

2 1 10 0.0050 1.5000 2.5000 ! C-Atom
2 1 12 0.0050 20.0000 70.0000
2 1 20 0.0050 0.0001 25.0000
2 1 21 0.0050 0.0001 40.0000
2 1 22 0.0050 0.0001 25.0000
2 1 25 0.0050 -20.0000 2.0000
2 1 26 0.0050 2.0000 4.5000
2 1 29 0.0050 2.7000 3.1000
2 2 10 0.0050 4.0000 6.0000 ! H-Atom
2 2 20 0.0050 0.0001 5.0000
2 2 21 0.0050 0.0001 5.0000
2 2 22 0.0050 0.0001 5.0000
2 2 25 0.0050 -20.0000 2.0000
2 4 1 0.0050 1.6000 1.8500 ! S-Atom
2 4 4 0.0050 1.7000 2.0000
2 4 5 0.0050 0.0001 1.0000
2 4 9 0.0050 8.0000 12.0000
2 4 10 0.0050 4.0000 6.0000
2 4 12 0.0050 20.0000 70.0000
2 4 20 0.0050 0.0001 25.0000
2 4 21 0.0050 0.0001 40.0000
2 4 22 0.0050 0.0001 25.0000
2 4 25 0.0050 -20.0000 2.0000
2 4 26 0.0050 2.0000 4.5000
2 4 29 0.0050 2.7000 3.1000
3 4 1 0.0050 121.0000 125.0000 ! CS-Bond
3 4 4 0.0050 -1.0000 2.0000
3 4 8 0.0050 0.0001 1.0000
3 4 9 0.0050 0.0001 10.0000
3 4 13 0.0050 -0.3000 -0.0001
3 4 14 0.0050 2.0000 10.0000
3 7 1 0.0050 120.0000 160.0000 ! SH-Bond
3 7 4 0.0050 -1.0000 2.0000

```

A. REAXFF Optimization Input Files

```
3 7 8 0.0050 0.0001 1.0000
3 7 9 0.0050 0.0001 10.0000
3 7 13 0.0050 -0.3000 -0.0001
3 7 14 0.0050 2.0000 10.0000
3 10 1 0.0050 116.0000 119.0000 ! SS-Bond
3 10 4 0.0050 -1.0000 2.0000
3 10 8 0.0050 0.0001 1.0000
3 10 9 0.0050 0.0001 10.0000
3 10 13 0.0050 -0.3000 -0.0001
3 10 14 0.0050 2.0000 10.0000
4 3 1 0.0050 0.0001 1.0000 ! CS-OD
4 3 2 0.0050 1.8000 2.2000
4 3 3 0.0050 8.0000 12.0000
4 3 4 0.0050 1.3500 1.5500
4 5 1 0.0050 0.0001 0.5000 ! SH-OD
4 5 2 0.0050 1.3000 1.6000
4 5 3 0.0050 8.0000 12.0000
4 5 4 0.0050 1.2000 1.4500
5 4 1 0.0050 50.0000 95.0000 ! CCS-Angle
5 4 2 0.0050 0.0001 50.0000
5 4 3 0.0050 0.0001 10.0000
5 4 5 0.0050 0.0001 5.0000
5 4 7 0.0050 1.0000 3.0000
5 10 1 0.0050 50.0000 95.0000 ! CSC-Angle
5 10 2 0.0050 0.0001 50.0000
5 10 3 0.0050 0.0001 10.0000
5 10 5 0.0050 0.0001 5.0000
5 10 7 0.0050 1.0000 3.0000
5 13 1 0.0050 50.0000 95.0000 ! CSS-Angle
5 13 2 0.0050 0.0001 50.0000
5 13 3 0.0050 0.0001 10.0000
5 13 5 0.0050 0.0001 5.0000
5 13 7 0.0050 1.0000 3.0000
5 20 1 0.0050 50.0000 95.0000 ! HSH-Angle
5 20 2 0.0050 0.0001 50.0000
5 20 3 0.0050 0.0001 5.0000
5 20 5 0.0050 1.0000 3.0000
5 22 1 0.0050 50.0000 95.0000 ! HSS-Angle
5 22 2 0.0050 0.0001 50.0000
5 22 3 0.0050 0.0001 5.0000
5 22 5 0.0050 1.0000 3.0000
5 29 1 0.0050 50.0000 95.0000 ! SCS-Angle
```

A. REAXFF Optimization Input Files

```
5 29 2 0.0050 0.0001 50.0000
5 29 3 0.0050 0.0001 10.0000
5 29 5 0.0050 0.0001 5.0000
5 29 7 0.0050 1.0000 3.0000
6 5 1 0.0050 0.0001 15.0000 ! CSSC-Dihedral
6 5 2 0.0050 -20.0000 -12.0000
6 5 3 0.0050 -3.0000 3.0000
6 5 4 0.0050 0.0000 -10.0000
6 15 1 0.0050 0.0001 15.0000 ! HSSH-Dihedral
6 15 2 0.0050 -22.0000 -10.0000
6 15 3 0.0050 -8.0000 8.0000
6 15 4 0.0050 0.0000 -10.0000
```

A.5. The sulfur.ogo File

The input file used for the optimization sets up the evolutionary algorithm. For an explanation of all input parameter please refer to the OGOLEM manual^[67].

```
###OGOLEM###
AlternativeInput=true
MaxIterLocOpt=50
<ADAPTIVE>
# ReaxFF initialisation stuff
AdaptivableChoice=ReaxFF:80,80,80,90,90,90;LocOptAlgo=lbfgs:ReaxFF:80,80,80,90,90,
90^7^0.9^50^2
FitnessCalculation=squared
ConvertWeights=true
AlternativeInput=true
ImmediateReturnInFitCalc=true
#ParamSeedingFolder=seed

#Globopt stuff
#ParameterGlobOpt=parameters{xover(portugal:nocuts=1)mutation(arctic:mode=multi,
submode=current,shape=0.01)}
#ParameterGlobOpt=parameters{xover(multiple:20%arctic:nomix=4,order=2,random|
80%portugal:nocuts=3)mutation(multiple:20%germany:mode=all|80%arctic:mode=multi,
submode=current,shape=0.01)}
ParameterGlobOpt=parameters{xover(multiple:20%arctic:nomix=10,order=2,random|
80%portugal:nocuts=4)mutation(multiple:20%germany:mode=all|40%arctic:mode=multi,
submode=current,shape=0.01|40%arctic:mode=multi,submode=current,shape=0.8)}
PopulationSize=500
ParamGlobOptIter=2500000
```

A. REAXFF Optimization Input Files

```
CrossProbability=0.8
MutationRatio=0.2
ParentsChoice=fitnessrankbased:gausswidth=0.3
# ParentsChoice=fitnessvaluebased:gausswidth=0.1
# ParentsChoice=fitnessrankbased:gausswidth=0.4
# ParentsChoice=fitnessvaluebased:gausswidth=0.4

#Niching stuff
# DoNiching=false
DoNiching=true
NichesPerDim=30
MaxIndividualsPerNiche=2
# WhichNicher:2;10
# WhichNicher:2;20
WhichNicher=2;75
# WhichNicher:2;80

#Deactivate Diversity Check and locopt
ParamThreshDiv=0
ParamFitnessDiv=0
AnyParamLocOpt=false

#Set output level
ParamsToSerial=50000
ParamSerializeAfterNewBest=false
IncrementParamPool=true
NoBestCountsToFlush=90000000
ParameterDetailedStats=false
WriteEveryParameterSet=false
ParamBorderPrint=false
AnyParamHistory=false
NicheStats=90000000
HistoryRecordsToASCII=10000000
HistoryRecordsToSerial=10000000

## max tasks to submit
MaxTasksToSubmit=1000

</ADAPTIVE>
```


B. Setting up `params` files

In the detailed discussion of the `params` file in section 4.4.1 much advice was given as to how to construct `params` files for a successful optimization. All this advice is summarized here as a quick reference.

Table B.1.: Quick reference for setting up bonding parameters in `params` files for global optimization. The parameters are identified by their name as it is used in the supporting information to the publication by *Mueller, van Duin* and *Goddard* from 2010^[109] and their three integer coordinates to locate them in the `ffield` file. The “X” is a wildcard to make clear that any parameter set of the specified block can hold these parameters. Reasonable boundaries are given in the columns p_{\min} and p_{\max} . When no general recommendation can be made exemplary values for the carbon atom and carbon carbon bond respectively are given in parenthesis. The last column features more general constraints and remarks to shrink the search space to the necessary minimum while retaining the most flexibility of REAXFF.

Parameter	ffield coordinate	p_{\min}	p_{\max}	Additional constraints
r_0^σ	2 X 1, 4 X 4	- (1.30)	- (1.45)	$> r_0^\pi, > r_0^{\pi\pi}, < r_{single}^{eq}$
r_0^π	2 X 7, 4 X 5	- (1.22)	- (1.30)	$< r_0^\sigma, > r_0^{\pi\pi}, < r_{double}^{eq}$
$r_0^{\pi\pi}$	2 X 17, 4 X 6	- (1.17)	- (1.22)	$< r_0^\sigma, < r_0^{\pi\pi}, \approx r_{double}^{eq}$
p_{bo1}	3 X 13	-0.3	0.0	$< p_{bo3}, < p_{bo5}$
p_{bo2}	3 X 14	2	10	$< p_{bo4}, < p_{bo6}$
p_{bo3}	3 X 10	-0.5	0.0	$> p_{bo1}, < p_{bo5}$
p_{bo4}	3 X 11	5	20	$> p_{bo2}, < p_{bo6}$
p_{bo5}	3 X 5	-0.7	0.0	$> p_{bo1}, > p_{bo3}$
p_{bo6}	3 X 7	8	30	$> p_{bo4}, > p_{bo6}$
p_{be1}	3 X 4	-1.2	1.0	consider existing ffield
p_{be2}	3 X 9	0.01	20.0	$< 1/a$ if $a > 0$, consider existing ffield
bo_{131}	2 X 21	- (0.0)	- (40.0)	Depending on the element
bo_{132}	2 X 20	- (0.0)	- (40.0)	Depending on the element
bo_{133}	2 X 22	- (0.0)	- (40.0)	Depending on the element
D_e^σ	3 X 1	$\approx 1.2 \cdot \text{BDE}$	$\approx 1.65 \cdot \text{BDE}$	Increment from single bond
D_e^π	3 X 2	$\approx 1.2 \cdot \text{BDE}$	$\approx 1.65 \cdot \text{BDE}$	Increment from double bond
$D_e^{\pi\pi}$	3 X 3	$\approx 1.2 \cdot \text{BDE}$	$\approx 1.65 \cdot \text{BDE}$	Increment from double bond

Table B.2.: Quick reference for setting up angle and torsional parameters in `params` files for global optimization. The parameters are identified by their name as it is used in the supporting information to the publication by *Mueller, van Duin and Goddard* from 2010^[109] and their three integer coordinates to locate them in the `ffield` file. The “X” is a wildcard to make clear that any parameter set of the specified block can hold these parameters. Reasonable boundaries are given in the columns p_{\min} and p_{\max} . The last column features more general constraints and remarks to shrink the search space to the necessary minimum while retaining the most flexibility of REAXFF.

Parameter	ffield coordinate	p_{\min}	p_{\max}	Additional constraints
p_{val1}	5 X 2	$2 \cdot E_{\text{barrier}} - 5$	$2 \cdot E_{\text{barrier}} + 5$	Mind force constants
p_{val2}	5 X 3	0.5	10.0	Mind force constants
p_{val3}	2 X 26	2.0	4.0	
p_{val4}	5 X 7	1.0	4.0	
p_{val5}	2 X 29	2.0	4.0	
p_{val7}	5 X 5	0.0	3.0	
$\theta_{0,0}$	5 X 1	60.0	80.0	50 – 90 for more extreme geometries
V_1	6 X 1	-100.0	100.0	Get start values and sample $\pm 5 - 10$
V_2	6 X 2	-100.0	100.0	Get start values and sample $\pm 5 - 10$
V_3	6 X 3	-100.0	100.0	Get start values and sample $\pm 5 - 10$
p_{tor1}	6 X 4	-1.0	-10.0	

B. Setting up *params* files

Table B.3.: Quick reference for setting up nonbonding parameters in *params* files for global optimization. The parameters are identified by their name as it is used in the supporting information to the publication by *Mueller, van Duin* and *Goddard* from 2010^[109] and their three integer coordinates to locate them in the `ffield` file. The “X” is a wildcard to make clear that any parameter set of the specified block can hold these parameters. Reasonable boundaries are given in the columns p_{\min} and p_{\max} . When no general recommendation can be made exemplary values for the carbon atom and carbon carbon bond respectively are given in parenthesis. The last column features more general constraints and remarks to shrink the search space to the necessary minimum while retaining the most flexibility of REAXFF.

Parameter	ffield coordinate	p_{\min}	p_{\max}	Additional constraints
r_{vdw}	2 X 4, 4 X 1	(1.8)	(2.0)	Ele. or bond specific start value
D_e	2 X 5, 4 X 2	0.0	1.0	
α	2 X 9, 4 X 3	5	15	
γ_{vdw}	2 X 10	(2.0)	(2.3)	Ele. specific start value
χ	2 X 14	(5.5)	(6.0)	Ele. specific start value
η	2 X 15	(6.75)	(7.25)	Ele. specific start value
γ	2 X 6	0.7	1.2	Ele. specific start value

C. Triazole Reference Geometries

This section contains an overview of the 57 final geometries used for the optimization of the 1,2,3-triazole force field by *Steffen*. The reference geometries are grouped in three categories. The 1,2,3-triazole educt structures (fig. C.1), alkyne product structures (fig. C.2) and azide product structures (fig. C.3).

C. Triazole Reference Geometries

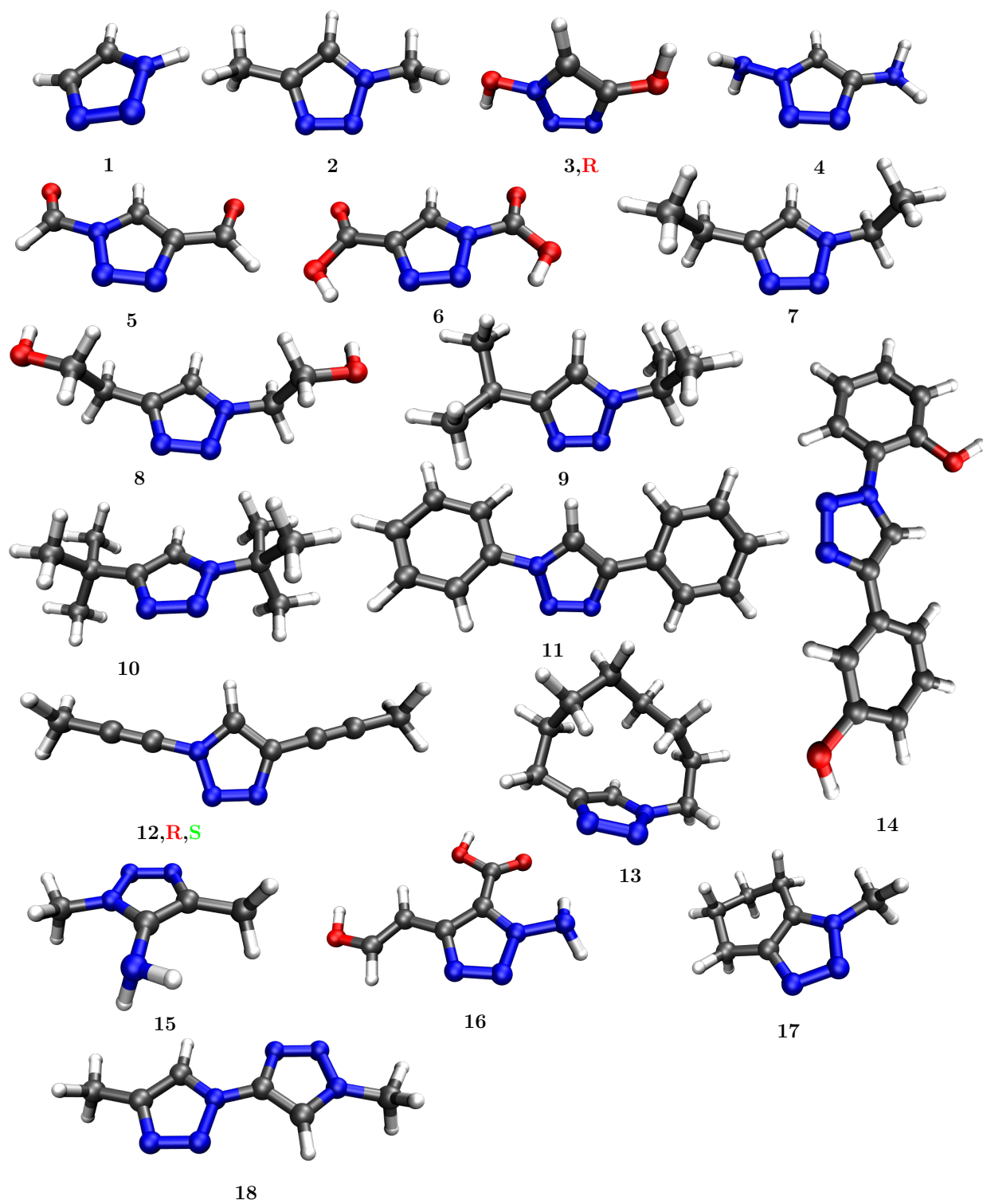


Figure C.1.: Set of 1,2,3-triazole reference structures used in *Steffen's* optimization. The structures were used for geometry and charge references. The figure is reprinted from *Steffens* F3 practical course thesis^[4].

C. Triazole Reference Geometries

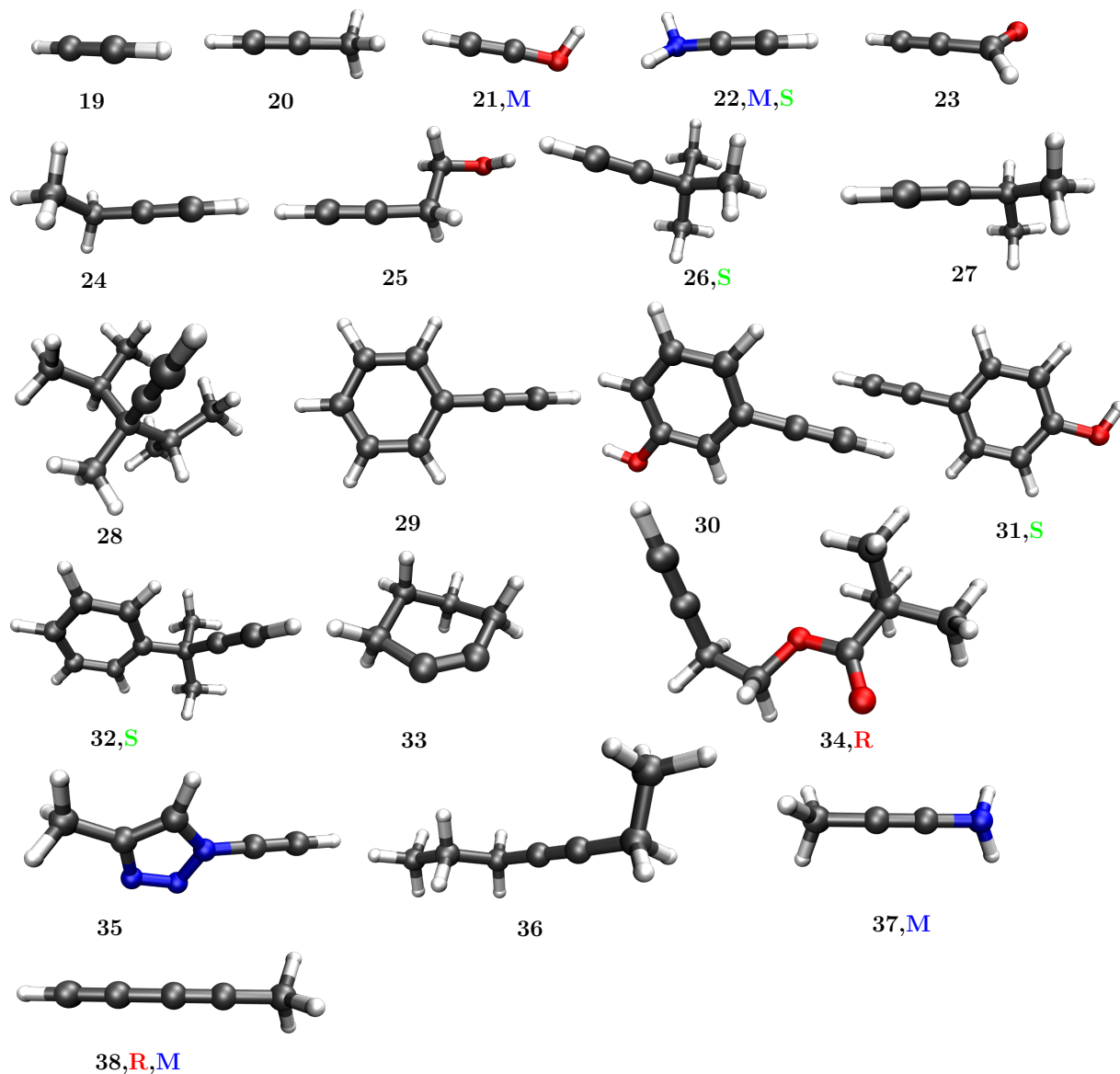


Figure C.2.: Set of alkyne reference structures used in *Steffen's* optimization. The structures were used for geometry and charge references. The figure is reprinted from *Steffens* F3 practical course thesis^[4].

C. Triazole Reference Geometries

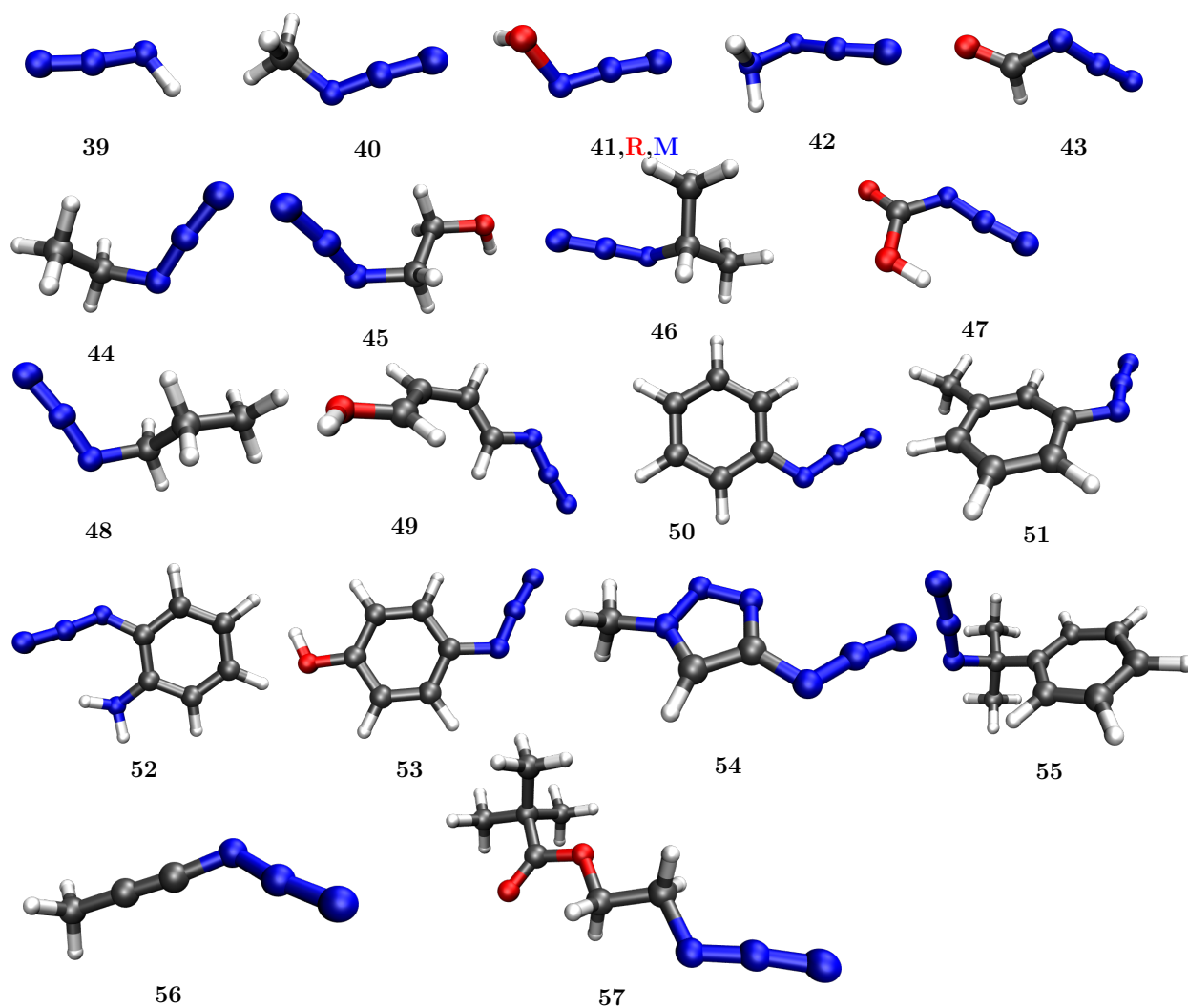


Figure C.3.: Set of azide reference structures used in *Steffen's* optimization. The structures were used for geometry and charge references. The figure is reprinted from *Steffens* F3 practical course thesis^[4].