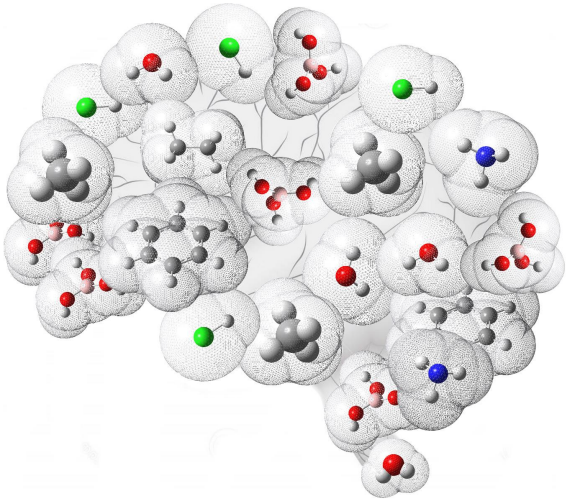
Developing Advanced Theoretical Methods For Enhanced Study of Thermochemistry



Dissertation

in partial fulfillment of the requirements for the degree

Dr. rer. nat.

of the Faculty of Mathematics and Natural Sciences at Kiel University

submitted by

Amin Alibakhshi

Kiel, September 2021

Developing advanced theoretical methods for enhanced study of thermochemistry

Dissertation, submitted by Amin Alibakhshi

Kiel, September 2021

First referee:	Prof. Dr. Bernd Hartke
Second referee:	Prof. Dr. Gernot Friedrichs
Date of the oral examination:	25.10.2021
Approved for publication:	25.10.2021

Abstract

Theoretical evaluation of thermochemistry is one of the most extensively required applications of theoretical chemistry in numerous cutting-edge scientific fields and technologies. The present thesis aims at developing advanced theoretical methods for enhanced evaluation of thermochemistry and verifying their performance through challenging case studies. To that end, we employ a wide range of tools and methods including static ab-initio calculations, classic and path integral molecular dynamics simulations, thermodynamic modeling and versatile machine learning approaches.

Considering that the liquid phase is the most challenging state of matter from the computational thermochemistry point of view, a special focus of the present thesis is on developing methods that allow a more rigorous study of this state. We achieve this goal partly by directly calculating liquid phase thermochemistry via ab-initio computations and path integral molecular dynamics simulations. Additionally, we also develop methods that address the liquid phase indirectly, through commonly employed thermodynamic cycles that require less challenging computations for the gas phase and phase change processes. For that purpose, we provide efficient models for a reliable and accurate evaluation of phase change thermodynamics via vaporization enthalpy and solvation free energy, based on statistical thermodynamics modeling as well as machine learning. Considering the importance of appropriately treating solvent effects in studying liquid phase thermochemistry, improvements of the two common approaches available for this purpose, namely the explicit and implicit solvent methods, are the other main focuses of the present thesis. For the explicit solvent method, we propose a cost-effective approach for multi-configuration ab-initio computation and partial normal mode analysis and employing a Boltzmann-weighted averaging for a rigorous integration of the results. For the implicit solvent approach, we propose thermodynamically effective molecular surfaces as well as machine learning models for achieving significantly better performance.

We demonstrate the efficiency of the developed methods for a diverse range of applications, including high-precision theoretical evaluation of combustion enthalpy, equilibrium constants of isotope exchange reactions, flash points of pure hydrocarbons, solvation free energy, vaporization enthalpy, potential energy surfaces, and inhibition of a cytochrome P450 enzyme in the human body by drug candidates.

Kurzzusammenfassung

Die theoretische Auswertung von Thermochemie ist eine der am weitesten verbreiteten Anwendungen von theoretischer Chemie in vielen hochaktuellen Wissenschaftsgebieten und Technologien. Die vorliegende Arbeit hat sich zum Ziel gesetzt, fortgeschrittene theoretische Methoden zur verbesserten Auswertung von Thermochemie zu entwickeln und deren Fähigkeiten durch anspruchsvolle Fallstudien zu verifizieren. Dazu wurde im Rahmen dieser Arbeit eine breite Auswahl von Werkzeugen und Methoden genutzt, darunter ab-initio Rechnungen, klassische und Pfadintegral Moleküldynamiksimulationen, thermodynamische Modellierungen und viele Ansätze aus dem Bereich des maschinellen Lernens. Eingedenk der Tatsache, dass die flüssige Phase vom Standpunkt der Computer Thermochemie betrachtet den anspruchsvollsten Aggregatszustand der Materie darstellt, liegt der Schwerpunkt dieser Arbeit auf der Entwicklung von Methoden, die eine präzisere Beschreibung von Flüssigkeiten ermöglichen. Dieses Ziel wird zum einen durch die direkte Berechnung von thermochemischen Parametern der flüssigen Phase auf Basis von ab-initio-Rechnungen und Pfadintegral- Moleküldynamiksimulationen erreicht. Zum anderen wurden Methoden entwickelt, welche die flüssige Phase indirekt adressieren, namentlich durch routinemäßig angewandte thermodynamische Zyklen, die mit weniger aufwändigeren Berechnungen der Gasphase und von Phasenübergangsprozessen auskommen. Zu diesem Zweck werden in dieser Arbeit effiziente Modelle zur akkuraten und zuverlässigen Beschreibung der Phasenübergangsthermodynamik basierend auf Verdampfungsenthalpien und Freien Lösungsenergien präsentiert, die auf Methoden der statistischen Thermodynamik und maschinellen Lernens aufbauen. Angesichts der zentralen Rolle, die der Beschreibung von Lösungsmitteleffekten im Studium der Flüssigphasenthermochemie zukommt, stellen Verbesserungen von zwei Ansätzen zur Beschreibung dieser Effekte, der expliziten und der impliziten Löungsmittelmethode, die weiteren Schwerpunkte der vorliegenden Arbeit dar. Für die explizite Lösungsmittelmethode wird ein kostengünstiger Ansatz für die Multikonfigurations-ab-initio Berechnung und partielle Normalmodenanalyse vorgeschlagen, deren Resultate durch eine Boltzmann-gewichtete Mittelung zu einer präzisen Integration der Ergebnisse eingesetzt werden. Für die implizite Lösungsmittelmethode wird die Nutzung von thermodynamisch effektiven molekularen Oberflächen und maschinellen Lernmethoden für eine deutlich bessere Performanz vorgeschlagen. Die Effektivität der entwickelten Methoden wird anhand einer Reihe von Anwendungen demonstriert: hochpräzise theoretische Auswertungen von Verbrennungsenthalpien, der Berechnung von Gleichgewichtskonstanten von Isotopenaustauschreaktionen, Siedepunkten von reinen Kohlenwasserstoffen, freien Lösungsenthalpien, Verdampfungsenthalpien und potentiellen Energieflächen sowie der Simulation der Inihibition des Cytochrome P450 Enzyms im menschlichen Körper durch verschiedene Wirkstoffe.

Acknowledgements

Completing this PhD would have not been possible without the kind helps and supports of many individuals.
I would like to thank my PhD supervisor, Prof. Dr. Bernd Hartke, for providing the opportunity to doing my PhD in his research group.
I am also deeply thankful to Dr. Jan Fietzke, one of my previous supervisors in the Geomar Helmholtz center for ocean research Kiel, for all his kind supports, and for providing me the freedom to follow up and enjoy my main scientific interests.
I am thankful to my colleagues in the research group of Prof. Hartke, for the always enjoyable and relaxing time I had there and for being so supportive whenever I needed their guidance. Specially, I am thankful to Julien Steffen, for many interesting scientific discussions and for his efforts in tailoring his excellent software EVB-QMDFF, which he recently developed as one of the several outstanding outcomes of his PhD, based on my needs.
I am thankful to Prof. Dr. Gernot Friedrichs for accepting to be the second referee of my thesis and his careful review of my works.
Every scientific achievement I have had so far or will have in the future is indebted to my high school physics teacher in the Farhang high school, Mr. Almasieh. With his patience, dedication, and love to physics, he could draw a beautiful picture of physics and mathematics in my mind which could alter my impression not only about those fields but also science forever. I am also thankful to Prof. Dr. Hamid Modarress in the Amirkabir University of Technology, who introduced me to the fundamentals of statistical mechanics, quantum mechanics, and molecular simulation and made the most abstract concepts in those fields digestible for me. The knowledge I gained in his astounding lectures was definitely one of the main privileges which enabled me to successfully complete my PhD.
I am thankful to Dr. Simone Knief and Dr. Carsten Balzer in high performance computing center of Kiel University for their supports and assistance in running my scientific computations there.
Last but not least, I am thankful to my beloved family, for all their endless love, passion and supports throughout my scientific and personal life.

This PhD thesis has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643084.

Table of contents

List of figures

List of tables

Nomenclature

Roman Symbols

A	Helmholtz free energy
G	Gibbs free energy
g_i	degeneracy of the state i
I	principal moment of inertia
k_B	Boltzmann constant
K_{eq}	equilibrium constant
h	Planck constant
P_c	critical pressure
Q	partition function
R	universal gas constant
R_g	radius of gyrration
S	entropy
T	temperature
T_c	critical temperature
U	potential energy
P_{sat}	saturation vapor pressure
V	molar volume

Greek Symbols

β $1/k_B T$

ε_i energy of the state i

$\varepsilon_{elec.}$ electronic energy state

$\varepsilon_{rot.}$ rotational energy state

$\varepsilon_{trans.}$ translational energy state

$\varepsilon_{vib.}$ vibrational energy state

ϕ electric potential

ΔG_{solv} solvation free energy

ΔH enthalpy of reaction

ΔH_{vap} vaporization enthalpy

$\Omega(N,V,E)$ total number of microstates in microcanonical ensemble

ω_j angular frequency for the jth mode of vibration

ρ charge distribution

Acronyms / Abbreviations

AAD average absolute deviation

AARE average absolute relative error

ANN artificial neural networks

NMA normal mode analysis

MD molecular dynamics

ML machine learning

MUE mean unsigned error

NBP normal boiling point

PIMD path integral molecular dynamics

RPF reduced partition function

RRHO rigid-rotor harmonic oscillator approximation

SAS solvent-accessible surfaces

SES solvent excluded surfaces

“Living is worthwhile if one can contribute in some small way to this endless chain of progress.”

Paul A.M. Dirac

Chapter 1

Introduction

1.1 Theoretical evaluation of thermochemistry: importance and applications

Theoretical evaluation of thermochemistry is one of the most widely required applications of theoretical and computational chemistry in many cutting-edge scientific and technological advancements. As can be implied by its definition, thermochemistry mainly deals with employing fundamental thermodynamic quantities, specifically enthalpy, free energy, and entropy, to determine major aspects of chemical reactions, such as the possibility of reactions occurring, their equilibrium constants, and rates [1,2]. In life science, studying thermochemistry plays a key role in evaluating the tendency of a drug to attach to the active site of an enzyme via computing the free energy or entropy of docking and has contributed to developing smarter and more effective medicines [3-7]. Similarly, theoretical thermochemistry has allowed unraveling the pathways and mechanisms of many diseases [8-12], protein functions [13-16], or viral activities [17-20]. As the most tangible example of the latter application for those who read this dissertation in 2021 and probably several years after that, we can refer to theoretical attempts to unravel the activity mechanisms of coronaviruses and exploit them to suppress the COVID-19 pandemic [21-24]. As examples of employing thermochemistry in cutting-edge technological applications, we can refer to its role in designing advanced and more efficient batteries [25-28], nanomaterials [29-32], and many other applications.

The theoretical study of thermochemistry can be generally regarded as one of the two major and indispensable components of computational and theoretical chemistry. While quantum chemistry, as one of the main cornerstones, mainly deals with theoretical chemistry at zero kelvin, thermochemistry complements it by extending the domain of application to higher temperatures where the majority of real-life chemistry occurs.

The modern theoretical thermochemistry in precursory form, to the author's opinion, began with the pioneering works of James Clerk Maxwell and Ludwig Boltzmann in formulating the Maxwell-Boltzmann statistics. Their ground-breaking finding resulted in a breakthrough not only in theoretical thermochemistry but also in modern science, due to establishing the building blocks of statistical thermodynamics as the keystone of theoretical thermochemistry as well as justifying the reality of atoms and molecules [33].

Following the early works of Maxwell and Boltzmann, the majority of significant scientific achievements in theoretical thermochemistry for several decades afterward have been mainly limited either to low-density gases or solid states, for similar reasons: For low density-gases, we can treat the atoms or molecules as individual particles in vacuo and therefore ignore their interactions with other particles. For the solid-state, we deal with organized structures where atoms can only vibrate around specific sites, which makes their theoretical study tractable.

For the liquid phase, however, the chaotic motion of atoms and molecules, which results in a quite diverse range of configurations, inter-atomic interactions, and thermodynamic quantities, substantially adds to the scientific challenges and causes most of the theoretical approaches applicable for the gas or solid states to break down for the liquid phase. For this reason, theoretical studies of thermochemistry for the liquid phase are commonly carried out via computer simulation techniques [34]. And for the same reason, the majority of advancements in studying the thermochemistry of liquids have become possible mainly in the past few decades and by the development of computer simulation algorithms as well as rapid growth of computational facilities.

With this motivation, the present study aims to develop advanced methods contributing to improve theoretical evaluation of thermochemistry with a specific focus on the liquid phase. To that end, we will employ conventional computer algorithms and techniques as well as analytical and machine learning models for a rigorous and straightforward evaluation of thermodynamic quantities in solution. Our other main focus will be to study thermochemistry associated with phase change through the thermodynamic cycle depicted in figure 1.1 as a commonly employed approach for relieving the computational challenges in studying solution thermodynamics [35-39]. The main reason behind this is that similar to experimental measurements, theoretical evaluation of the changes in thermodynamic quantities for a process is far more convenient and less error-prone compared to evaluating absolute values for a single state. Consequently, quantities like vaporization enthalpy or solvation free energy, which are the changes in enthalpy and free energy for phase change from liquid to the gas phase and from gas to the liquid phase, respectively, are much more frequently encountered in

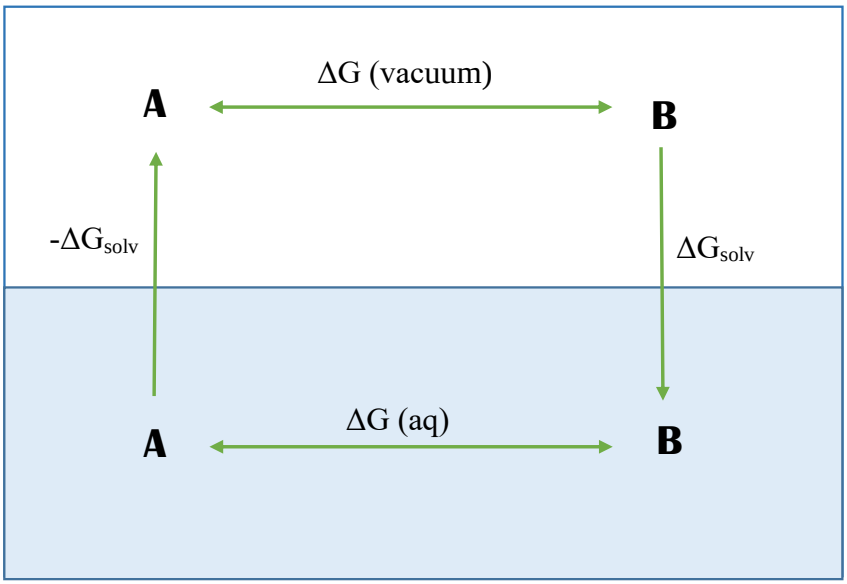


Fig. 1.1 Schematic demonstration of evaluating the free energy of reaction in solution via thermodynamic cycles. Through the illustrated thermodynamic cycle, the thermodynamic quantities for the lower leg of this cycle are indirectly obtained using less challenging evaluation of the in vacuo and phase change free energies.

thermochemistry studies than the absolute values of enthalpy or free energy of a single-phase [40].

In addition to offering a much lower computational challenge, once the thermodynamic quantities associated with phase change are found, they can then be used for a convenient evaluation of absolute values of those thermodynamic quantities for the liquid phase. That only requires subtracting the more conveniently obtainable absolute values of the thermodynamic quantities in the gas phase from those obtained for the phase change process.

Among potential thermodynamic quantities associated with phase change, our focus will mainly be on solvation free energy, which will be studied in chapters 8 and 9, and vaporization enthalpy, which will be the topic of study in chapters 5 and 7. One of the main reasons behind choosing these two quantities to evaluate phase change thermodynamics stems from the existence of well-established and extensive datasets of experimentally determined data for them, which are more readily available compared to entropy or the internal energy of the phase change. Additionally, by determining the vaporization enthalpy and solvation free energy, we can evaluate all other thermodynamic quantities of the solution using fundamental thermodynamic relationships.

In the carried out researches presented throughout this thesis, the studied problems and employed theoretical methods and tools span a diverse range of scientific disciplines, including but not limited to theoretical chemistry, classical and statistical thermodynamics, quantum physics, artificial intelligence and computer science, biophysics, and geochemistry, as summarized in table 1.1.

Having the experience of academic education in multiple and diverse disciplines namely chemistry, chemical engineering, nanotechnology, geobiochemistry, and theoretical chemistry, allowed me to be able to work in the interphase of those disciplines and develop methods that can connect many of them together and fill the gaps in between. However, considering that many readers of the present dissertation probably will not have such diverse backgrounds, I provide a brief introduction to the theory and methods employed throughout this dissertation in chapter 2. One of my main attempts in that chapter is to avoid giving too specialized details and intricate mathematical formulations as they can be found in numerous available textbooks, some of which suggested for further readings in each section, as well as the provided details on specific methods in the methods section of each chapter. Accordingly, the main focus of chapter 2 is to present a quick and brief introduction to the theories and methods employed in this thesis, in a simple and understandable way for readers without a background in those subjects, as well as discussing the general idea behind their applications in this thesis.

1.2 Thesis organization

After introducing the employed methods and tools in chapter 2, we start our journey by studying the thermochemistry of combustion reactions in chapter 3, as one of the earliest and classic examples of employing theoretical thermochemistry for a practical application. This provides us with the possibility to benchmark the major conventional and classic methods in theoretical thermochemistry for the less challenging case of reactions in the gas phase.

Among the potential gas-phase reactions which can be used for this purpose, we consider the combustion reactions due to the availability of highly accurate benchmark sets for these reactions as well as their extensive scientific and technological applications. The other main motivation of studying combustion reactions stems from several reports of observing a remarkable inconsistency between theoretically predicted and experimentally measured data for them, in the literature [41-43].

In chapter 4, we employ methods and knowledge acquired through studying the gas-phase reactions in chapter 3 for studying the more challenging case of isotope exchange reactions in the solution phase. To that end, we evaluate the equilibrium constant of the

Table 1.1 Multidisciplinary aspects of the carried out researches

Scientific discipline	**Studied methods and applications**
Theoretical chemistry	Quantum mechanical evaluation of interatomic interactions and normal modes in chapters 3 and 4, computation of implicitly perturbed Hamiltonian via continuum solvation models in chapters 8 and 9
Statistical thermodynamics	Calculation of heat and equilibrium constant of reactions via partition functions in chapters 3 and 4, deriving a model for prediction of vaporization enthalpy in chapter 5
Classical thermodynamics	Evaluation of phase change thermodynamics impacts on combustion enthalpy in chapter 3, developing the physical model for predicting vaporization enthalpy and free energy of solvation in chapter 5
Quantum physics	Path integral molecular dynamics for evaluation of isotope effect in chapter 4
Artificial intelligence and computer science	The machine learning models and provided software codes in chapters 6-9
Geochemistry	Evaluation of isotope fractionation for important geochemical reactions in chapter 4

boron isotope exchange reaction between boric acid and borate for both pure and saline water, as one of the reactions with high geochemical importance [44,45]. In that chapter, we study and compare the two main approaches for considering solvent effects in solution phase thermochemistry, namely the implicit and explicit solvent approaches. The former approach and knowledge acquired through its application in chapter 4 is the cornerstone of proposing the thermodynamically effective molecular surfaces in chapter 5 and the more advanced continuum solvation methods introduced in chapters 8 and 9.

In chapters 3 and 4, one of the main employed theoretical approaches is the evaluation of the partition functions through Normal Mode Analysis (NMA) based on the rigid rotor harmonic oscillator approximation. This approach is not only one of the earliest but also one of the main tools to evaluate thermodynamic quantities from first principles and is the default method in almost all of the quantum chemistry software packages. Despite its widespread applications and convenience of usage, this approach suffers from several limitations such as neglecting the anharmonicity effects as well as configurational entropy and from difficulty in treating solvents explicitly. A robust and rapidly growing approach that can address all these limitations is the Path Integral Molecular Dynamics (PIMD) simulation [46]. With that motivation, in chapter 4, we also investigate the estimation of the equilibrium constant for the studied isotope exchange reaction via PIMD and develop methods that can improve numerical calculation of thermodynamic quantities in this approach.

In chapter 5, we study the predictability of vaporization enthalpy as another important and complementary thermodynamic quantity for fully characterizing thermochemistry in solution. To that end, by employing statistical thermodynamics of vaporization, we theoretically derive a physical model which allows us to accurately evaluate the vaporization enthalpy of various compounds from diverse chemical families and for a wide temperature range. Through this research, we also theoretically quantify the impact of molecular surfaces on solution thermodynamics, which can be of fundamental importance. In the context of this research, we propose the thermodynamically effective surfaces, as a more suitable alternative for conventionally accepted molecular surfaces like solvent excluded or van-der-Waals surfaces for studying thermochemistry. Further study of the vaporization enthalpy and the physical model proposed in this chapter will be the subject of chapter 7.

In all the research projects introduced so far, the main focus has been to study and develop theoretical methods which rely on exact physical rules governing the thermodynamics of solutions. Alongside these classical approaches, artificial intelligence and machine learning (ML) have recently emerged as a highly versatile and promising alternative for studying the most complicated challenges not only in theoretical chemistry but also in many other scientific fields. With that motivation, the second part of this thesis which includes the research works

presented in chapters 6 to 9, aims at exploiting machine learning for an enhanced study of thermochemistry in solution. For that purpose, the main ML tool employed in the present study will be Artificial Neural Networks (ANN), as one of the extremely powerful and most extensively applied tools for machine-learning function approximation. Accordingly, ANN is exploited to map the complicated dependency between potentially relevant quantum mechanically computed variables and the thermodynamic quantities of interest in solution.

Although a large number of available user-friendly software tools have made developing ANN models a convenient and straightforward task, still there are a number of intricacies that are commonly overlooked and can significantly affect the accuracy and reliability of ANN models. In chapter 6 we first benchmark such intricacies by studying the predictability of flash point based on the ANN and group contribution method for a large dataset. It allows us to select appropriate settings for the most important aspects of neural network models. The guidelines provided in that chapter will be exploited in the advanced machine learning models introduced for evaluation of vaporization enthalpy in chapter 7 and solvation free energy in chapters 8 and 9.

As the first example of employing ML for studying thermochemistry, in chapter 7 we introduce ML and algebraic models to extend the main achievements in the evaluation of vaporization enthalpy introduced in chapter 5.

As it was discussed earlier, one of the two main methods to study solvent effects in theoretical and computational chemistry is the implicit solvent approach. Due to its extensive applications in a very broad range of scientific fields, in chapter 8 we exploit machine learning to improve the accuracy of the widely applied implicit solvent models. To that end, ANN is applied for a more versatile integration of implicitly perturbed Hamiltonian energy attributes in conventional continuum solvation models. A further extension of this idea results in proposing the Implicitly Perturbed Hamiltonian (ImPerHam) in chapter 9, as a new class of general-purpose and highly efficient molecular representations for studying various scientific problems in molecular sciences.

We benchmark the efficiency of the ImPerHam representations in machine learning study of three challenging problems in molecular sciences which are inhibition of a cytochrome P450 enzyme in the human body by drug candidates, prediction of solvation free energy, and evaluation of molecular energies.

In this chapter, I briefly introduced the main objectives and goals of the present study. The connectivity of carried out research studies discussed above is schematically depicted in figure 1.2. Further details on individual projects are provided at the beginning of each chapter as well as in chapter 10 where we discuss the summary and outlook of this research project in more detail.

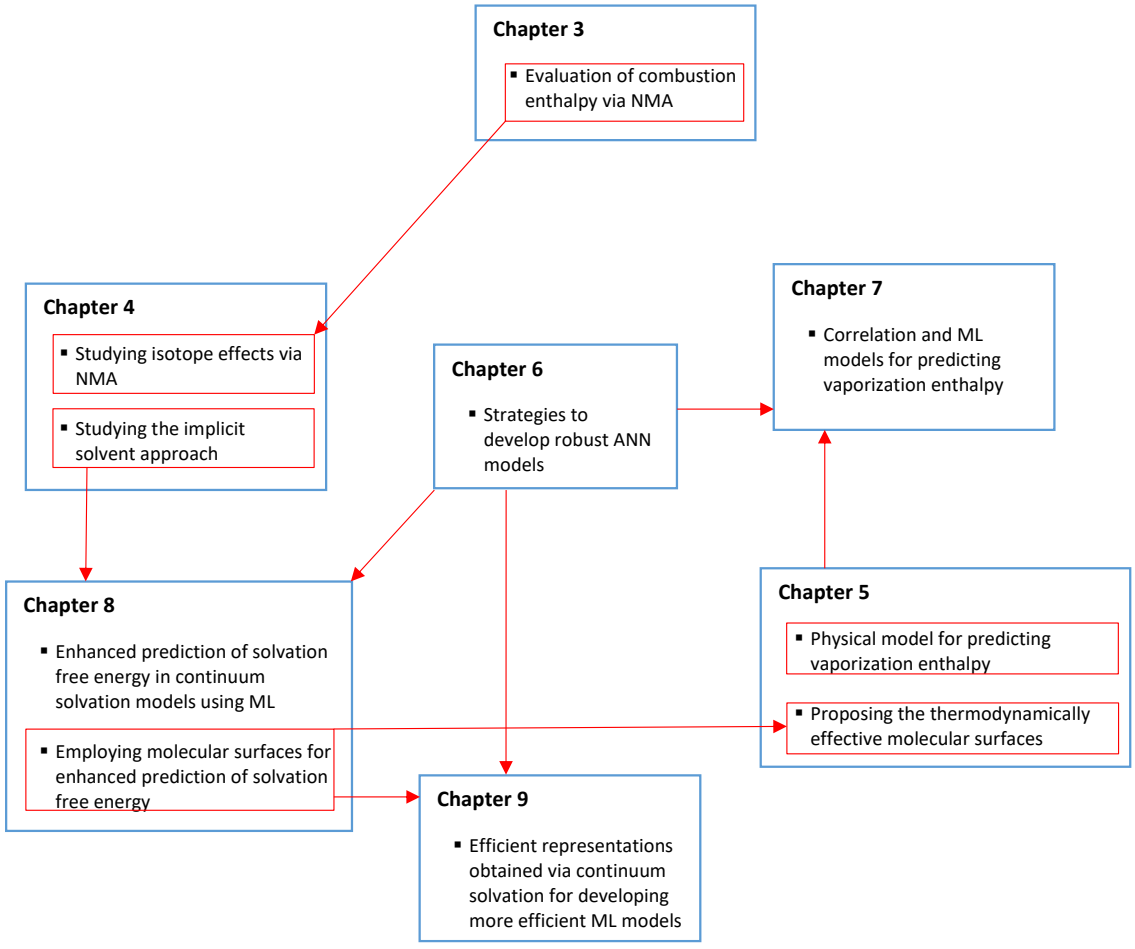


Fig. 1.2 Logic and connectivity between the subprojects of this thesis.

References:

1. Atkins, P.; De Paula, J., Elements of physical chemistry. Oxford University Press, USA: 2013.

2. Smith, J. M., Introduction to chemical engineering thermodynamics. ACS Publications: 1950.

3. Ahmed, B.; Khan, S.; Nouroz, F.; Farooq, U.; Khalid, S., Exploring multi-target inhibitors using in silico approach targeting cell cycle dysregulator–CDK proteins. Journal of Biomolecular Structure and Dynamics 2021, 1-15.

4. Kundu, D.; Dubey, V. K., Potential alternatives to current cholinesterase inhibitors: An in silico drug repurposing approach. Drug Development and Industrial Pharmacy 2021, (just-accepted), 1-30.

5. Su, K.-H.; Wu, C.-T.; Lin, S.-W.; Mori, S.; Liu, W.-M.; Yang, H.-C., Calculation of CYP450 protein–ligand binding and dissociation free energy paths. The Journal of Chemical Physics 2021, 155 (2), 025101.

6. Adeniyi, A. A.; Conradie, J., Computational insight into the anticholinesterase activities and electronic properties of physostigmine analogs. Future medicinal chemistry 2019, 11 (16), 1907-1928.

7. Magistrato, A.; Sgrignani, J.; Krause, R.; Cavalli, A., Single or multiple access channels to the CYP450s active site? An answer from free energy simulations of the human aromatase enzyme. The journal of physical chemistry letters 2017, 8 (9), 2036-2042.

8. Jing, J.; Tu, G.; Yu, H.; Huang, R.; Ming, X.; Zhan, H.; Zhan, F.; Xue, W., Copper (Cu 2+) ion-induced misfolding of tau protein R3 peptide revealed by enhanced molecular dynamics simulation. Physical Chemistry Chemical Physics 2021, 23 (20), 11717-11726.

9. Lee, K.-H.; Kuczera, K., Free energy simulations to understand the effect of Met→ Ala mutations at positions 205, 206 and 213 on stability of human prion protein. Biophysical Chemistry 2021, 275, 106620.

10. Serrano-Aparicio, N.; Moliner, V.; Swiderek, K., Nature of Irreversible Inhibition of Human 20S Proteasome by Salinosporamide A. The Critical Role of Lys–Asp Dyad Revealed from Electrostatic Effects Analysis. ACS Catalysis 2021, 11 (6), 3575-3589.

11. Leonard, C.; Phillips, C.; McCarty, J., Insight Into Seeded Tau Fibril Growth From Molecular Dynamics Simulation of the Alzheimer's Disease Protofibril Core. Frontiers in molecular biosciences 2021, 8, 109.

12. Aggarwal, L.; Biswas, P., Hydration Thermodynamics of the N-Terminal FAD Mutants of Amyloid-

β. Journal of Chemical Information and Modeling 2021, 61 (1), 298-310.

13. Katiyar, A.; Thompson, W. H., Temperature Dependence of Peptide Conformational Equilibria from Simulations at a Single Temperature. The Journal of Physical Chemistry A 2021, 125 (11), 2374-2384.

14. Li, J.; Hou, C.; Ma, X.; Guo, S.; Zhang, H.; Shi, L.; Liao, C.; Zheng, B.; Ye, L.; Yang, L., Entropy-Enthalpy Compensations Fold Proteins in Precise Ways. International Journal of Molecular Sciences 2021, 22 (17), 9653.

15. Akhter, N.; Qiao, W.; Shehu, A., An energy landscape treatment of decoy selection in template-free protein structure prediction. Computation 2018, 6 (2), 39.

16. Yang, C.; Jang, S.; Pak, Y., Computational Probing of Temperature-Dependent Unfolding of a Small Globular Protein: From Cold to Heat Denaturation. Journal of Chemical Theory and Computation 2020, 17 (1), 515-524.

17. Samandoulgou, I.; Fliss, I.; Jean, J., Adhesion of Norovirus to Surfaces: Contribution of Thermodynamic and Molecular Properties Using Virus-Like Particles. Food and Environmental Virology 2021, 1-12.

18. Rahman, M. M.; Biswas, S.; Islam, K. J.; Paul, A. S.; Mahato, S. K.; Ali, M. A.; Halim, M. A., Antiviral phytochemicals as potent inhibitors against NS3 protease of dengue virus. Computers in Biology and Medicine 2021, 134, 104492.

19. Popovic, M.; Minceva, M., A thermodynamic insight into viral infections: do viruses in a lytic cycle hijack cell metabolism due to their low Gibbs energy? Heliyon 2020, 6 (5), e03933.

20. Ali, N.; Khalil, R.; Nur-e-Alam, M.; Ahmed, S.; Ul-Haq, Z., Probing the mechanism of peptide binding to REV response element RNA of HIV-1; MD simulations and free energy calculations. Journal of Biomolecular Structure and Dynamics 2020, 1-10.

21. Pang, J.; Gao, S.; Sun, Z.; Yang, G., Discovery of small molecule PLpro inhibitor against COVID-19 using structure-based virtual screening, molecular dynamics simulation, and molecular mechanics/Generalized Born surface area (MM/GBSA) calculation. Structural chemistry 2021, 32 (2), 879-886.

22. Ding, H.-m.; Yin, Y.-w.; Sheng, Y.-j.; Ma, Y.-q., Accurate Evaluation on the Interactions of SARS-CoV-2 with Its Receptor ACE2 and Antibodies CR3022/CB6. Chinese Physics Letters 2021, 38 (1), 018701.

23. Liu, J.; Zhai, Y.; Liang, L.; Zhu, D.; Zhao, Q.; Qiu, Y., Molecular modeling evaluation of the binding effect of five protease inhibitors to COVID-19 main protease. Chemical Physics 2021, 542, 111080.

24. Li, Z.; Li, X.; Huang, Y.-Y.; Wu, Y.; Liu, R.; Zhou, L.; Lin, Y.; Wu, D.; Zhang, L.; Liu, H., Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual screening of existing drugs. Proceedings of the National Academy of Sciences 2020, 117 (44), 27381-

27387.

25. Lipka-Bartosik, P.; Mazurek, P.; Horodecki, M., Second law of thermodynamics for batteries with vacuum state. Quantum 2021, 5, 408.

26. Chen, Y.; Yang, X.; Luo, D.; Wen, R., Remaining available energy prediction for lithium-ion batteries considering electrothermal effect and energy conversion efficiency. Journal of Energy Storage 2021, 40, 102728.

27. Choi, Y.; Preston, T. J.; Adamczyk, A. J., Data-driven investigation of monosilane and ammonia co-pyrolysis to silicon-nitride-based ceramic nanomaterials. ChemPhysChem 2020, 21 (22), 2627-2642.

28. Lener, G.; Otero, M.; Barraco, D.; Leiva, E. P. M., Energetics of silica lithiation and its applications to lithium ion batteries. Electrochimica Acta 2018, 259, 1053-1058.

29. Agah, A.; Falahati, N., Studying Effect of Modifying Nano-Mineral Adsorbents on Efficiency of Dye Removal from Industrial Effluents. Journal of Mining and Environment 2021, 12 (1), 219-233.

30. Duan, H.; Cheng, Z.; Xue, Y.; Cui, Z.; Yang, M.; Wang, S., Influences of nano-effect on electrochemical thermodynamics of metal nanoparticles electrodes. Journal of Electroanalytical Chemistry 2021, 882, 115037.

31. Tong, W.-Y.; Zhao, T.-T.; Zhao, X.-F.; Wang, X.; Wu, Y.-B.; Yuan, C., Neutral nano-polygons with ultrashort Be–Be distances. Dalton Transactions 2019, 48 (42), 15802-15809.

32. El-Baba, T. J.; Clemmer, D. E., Solution thermochemistry of concanavalin A tetramer conformers measured by variable-temperature ESI-IMS-MS. International journal of mass spectrometry 2019, 443, 93-100.

33. Rosa, L. P.; Andrade, E.; Picciani, P.; Faber, J., Constructivism and Realism in Boltzmann's Thermodynamics' Atomism. Foundations of Physics 2020, 50 (11), 1270-1293.

34. Allen, M. P.; Tildesley, D. J., Computer simulation of liquids. Oxford university press: 2017.

35. Fowles, D. J.; Palmer, D. S.; Guo, R.; Price, S. L.; Mitchell, J. B., Toward Physics-Based Solubility Computation for Pharmaceuticals to Rival Informatics. Journal of chemical theory and computation 2021.

36. Itkis, D.; Cavallo, L.; Yashina, L. V.; Minenkov, Y., Ambiguities in solvation free energies from cluster-continuum quasichemical theory: lithium cation in protic and aprotic solvents. Physical Chemistry Chemical Physics 2021, 23 (30), 16077-16088.

37. Chen, R.; Deng, S.; Xu, W.; Zhao, L., A graphic analysis method of electrochemical systems for low-grade heat harvesting from a perspective of thermodynamic cycles. Energy 2020, 191, 116547.

38. Abraham, N. S.; Shirts, M. R., Statistical mechanical approximations to more efficiently determine polymorph free energy differences for small organic molecules. Journal of Chemical Theory and Computation 2020, 16 (10), 6503-6512.

39. Gapsys, V.; Seeliger, D.; de Groot, B. L., New soft-core potential function for molecular dynamics

based alchemical free energy calculations. Journal of chemical theory and computation 2012, 8 (7), 2373-2382.

40. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. Nature Communications 2021,12(1),1-7.

41. Mazzuca, J. W.; Downing, A. R.; Potter, C., Empirically corrected electronic structure calculations applied to the enthalpy of combustion physical chemistry laboratory. Journal of Chemical Education 2019, 96 (6), 1165-1170.

42. Audran, G.; Marque, S. R.; Siri, D.; Santelli, M., Enthalpy of Combustion on n-Alkanes. Quantum Chemical Calculations up to n-C60H122 and Power Law Distributions. ChemistrySelect 2018, 3 (31), 9113-9120.

43. Whyman, G.; Savoskin, M.; Yaroshenko, A.; Kapkan, L.; Popov, A., Straightforward ab initio calculation of enthalpies of combustion and formation of hydrocarbons. Journal of Molecular Structure: THEOCHEM 2003, 637 (1-3), 183-187.

44. Dickson, A. G.; Goyet, C. Handbook of methods for the analysis of the various parameters of the carbon dioxide system in sea water. Version 2; Oak Ridge National Lab., TN (United States): 1994.

45. Zeebe, R. E.; Sanyal, A.; Ortiz, J. D.; Wolf-Gladrow, D. A., A theoretical study of the kinetics of the boric acid–borate equilibrium in seawater. Marine Chemistry 2001, 73 (2), 113-124.

46. Cao, J.; Voth, G. A., The formulation of quantum statistical mechanics based on the Feynman path centroid density. I. Equilibrium properties. The Journal of chemical physics 1994, 100 (7), 5093-5105.

Chapter 2

Theory and background

This chapter provides an overview of the important theoretical concepts and methods employed throughout this thesis. It includes a brief introduction to statistical thermodynamics (section 2.1), evaluation of partition functions based on the rigid rotor harmonic oscillator approximation (section 2.2), normal mode analysis (section 2.3), classic and path integral molecular dynamics (sections 2.4 and 2.5), implicit solvent approach (section 2.6), and machine learning (section 2.7). The provided topics in the first four sections of this chapter are mainly the author's own approach in the presentation of the phenomenological topics. For the other sections, the provided subjects are mostly gathered from the sources cited within the text.

2.1 Statistical thermodynamics: a brief overview

The development of statistical thermodynamics as a scientific field was motivated by the inability of classical thermodynamics in providing a physical interpretation for fundamental thermodynamic quantities such as entropy and free energy and their connection with atomic-scale specifications of systems. In classical thermodynamics, entropy is defined as a hypothetic parameter to relate heat that is exchanged in a process to temperature [1], and free energy is also a Legendre transformation of internal energy to allow obtaining relationships as a function of experimentally measurable variables [2]. Consequently, classical thermodynamics has to rely on experimental measurements or empirical models to estimate those required fundamental quantities.

To address this limitation, in the late 18^{th} century Ludwig Boltzmann proposed a statistical interpretation for entropy, which later on became the foundation of statistical thermodynamics. Unlike the initially defined concepts, the statistical interpretation of fundamental quantities allows their direct evaluation via basic rules of physics and statistics. The groundbreaking

work of Boltzmann is undoubtedly one of the most eminent scientific achievements in the history of science and resulted in a breakthrough in many scientific fields.

Statistical thermodynamics follows very simple basics. It presumes that a system of particles[1] can be observed in numerous energy states. These energy states can result from different geometrical configurations. For example, different distances between the hydrogen atoms in a H_2 molecule result in different energies. Similarly, different speeds of particles can also result in different kinetic energies and therefore different energy states. Accordingly, for a system of N particles, the total energy of the system is a function of 3N coordinates and 3N conjugate momenta and all possible combinations of these 6N variables yield all possible energy states, as employed later in Eq. (2.8). Each possible energy state is called a microstate to specify its attribution to micro-scale details of the system.

The cornerstone of Boltzmann's work is based on the premise that each one of the possible energy states of a system is equally possible. It is very similar to tossing a dice that has numerous faces. For each tossing, the possibility of turning up each specific face is equally likely, but the possibility of observing specific numbers written on that face depends on how frequent that number is among all faces. For molecular systems, each configuration is similar to each face of a dice and the energy of that configuration is similar to the numbers written on the faces of the dice. Finding the probability of observing those energy states (the fraction of dice faces on which a specific value of energy is written) is the main goal in statistical thermodynamics. For that purpose, two different approaches, namely discrete and continuous formulations of statistical thermodynamics, are developed based on two different methodologies.

The discrete formulation, which is the earliest approach originally proposed by Boltzmann, is derived through considering the total number of ways N particles can be distributed over i groups (each group is in fact one energy state) which based on combinatorics is equal to [3]:

$$W = \frac{N! \prod_i {g_i}^{N_i}}{\prod_i N_i!}. \tag{2.1}$$

Here, g_i is the total number of distinguished states which have the same energy ε_i and is called the degeneracy of the state, and N_i is the total number of particles occupying the state with energy ε_i . Clearly, the most probable distribution of particles over energy states

[1] Here and throughout this thesis, by particle we actually mean Boltzmann particles, e.g. neutral atoms and closed shell molecules. Describing the statistical thermodynamics for the other types of particles i.e. bosons and fermions, is provided by Bose-Einstein and Fermi-Dirac statistics, respectively, and won't be considered in this thesis.

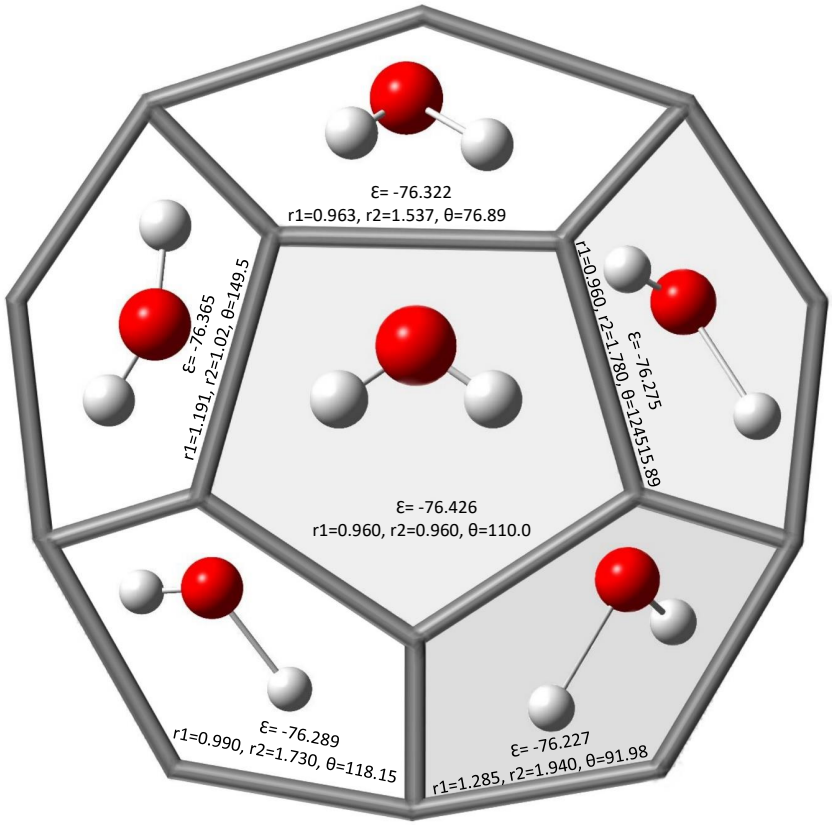


Fig. 2.1 Schematic illustration of Boltzmann statistics. For an imaginary water molecule with only 12 energy states depicted on the faces of this dice, the observed energy is obtained by a hypothetic experiment of tossing this dice numerous times every moment. For the real case, this dice is replaced by another one with many further faces. The fraction of faces with a specific value of energy is determined by the Boltzmann factor.
(For this demonstration, the energies are computed by B3LYP/6-311+G level of theory, the bond lengths in Angstrom and degrees are specified for each calculation)

is a distribution which maximizes W subjected to two constraints: The sum of particles distributed over all energy states based on the obtained distribution should be equal to the total number of particles in the system ($\sum_i N_i = N$), and the resulting total energy should also be equal to the energy of the isolated system ($\sum_i N_i \varepsilon_i = E$). As a classical method for solving optimization problems under constraint, the Lagrange method of undetermined multipliers can be employed to find the most probable distribution of particles, resulting in the so-called Maxwell-Boltzmann statistics stated as [3]:

$$\frac{N_i}{N} = \frac{g_i e^{-\frac{\varepsilon_i}{k_B T}}}{\sum_i g_i e^{-\frac{\varepsilon_i}{k_B T}}}, \tag{2.2}$$

where k_B is the Boltzmann constant and $\frac{N_i}{N}$ is the most probable fraction of particles that occupy a microstate with energy ε_i. Equivalently, $\frac{N_i}{N}$ is also equal to the probability of observing a microstate with energy ε_i in a particle. The reason is that if the probability of observing each energy state of a particle is denoted by P_i, for a very large number of particles at every moment, a fraction equal to P_i of them are observed at that energy state based on the ergodicity concept. Consequently, the ensemble average or expectation value of the total energy which is the statistical average of all energy states weighted by their respective probabilities is found as:

$$\langle E \rangle = \frac{\sum_i \varepsilon_i \; g_i e^{-\frac{\varepsilon_i}{k_B T}}}{\sum_i g_i e^{-\frac{\varepsilon_i}{k_B T}}} . \tag{2.3}$$

The denominator of Eq. (2.3) is called *partition function* and is commonly denoted by Q. As demonstrated in the following, partition functions play a central role in statistical thermodynamics, and not only the energy but also other fundamental quantities in thermodynamics are fully determined through them. For example, the obtained relationship for energy in Eq. (2.3) can be rewritten as:

$$\begin{aligned} \langle E \rangle = \frac{\partial \ln\left(\sum_i g_i e^{-\frac{\varepsilon_i}{k_B T}}\right)}{\partial\left(\frac{1}{k_B T}\right)} &= -\frac{\partial \ln Q}{\left(\frac{1}{k_B T}\right)^2 \partial (k_B T)} \\ &= -\frac{k_B T^2 \partial \ln Q}{\partial T} , \end{aligned} \tag{2.4}$$

where in the second line we inserted $d\left(\frac{1}{k_B T}\right) = -\left(\frac{1}{k_B T}\right)^2 d(k_B T)$ by the chain rule.

By substituting the found relationship for $\langle E \rangle$ introduced in Eq. (2.4) in the Gibbs-Helmholtz relationship defined as [1]:

$$\frac{d\left(\frac{\langle A \rangle}{T}\right)}{dT} = -\frac{\langle E \rangle}{T^2} , \tag{2.5}$$

we obtain the following relationship to define the Helmholtz free energy through partition functions:

$$\langle A \rangle = k_B T \; \ln Q . \tag{2.6}$$

Similarly, the other thermodynamic quantities such as entropy, enthalpy, and Gibbs free energy, which are related to internal and Helmholtz free energies via the fundamental variables (T, V, and P), can also be simply determined through the partition functions. For example, by substituting Eqs. (2.4) and (2.6) into the thermodynamic interpretation of Helmholtz free energy $\langle A\rangle = \langle E\rangle - TS$, entropy is found as:

$$S = \frac{-\langle A\rangle}{T} + \frac{\langle E\rangle}{T} = -k_B \ln Q + k_B T \frac{\partial \ln \mathrm{Q}}{\partial T}. \tag{2.7}$$

In addition to the discrete formulation of statistical thermodynamics introduced above, an alternative formulation which is known as the continuous formulation is also widely employed, especially in studying an ensemble of particles. According to the continuous formulation of statistical thermodynamics, for the microcanonical ensemble[2] with a total energy of E, via the Sturm–Liouville theory we find the total number of microstates as [4]:

$$\Omega(N,V,E) = \frac{1}{h^{3N}N!}\int\int_{V^N} \delta\left[H(q,p) - E\right]\, dqdp, \tag{2.8}$$

where h is the Planck constant, H is the Hamiltonian of the system, and q and p are the $3N$ positions and $3N$ momenta as discussed earlier, respectively. Via this evaluated number of microstates, the partition function for the canonical ensemble[3] can be obtained as [4]:

$$Q(N,V,T) = \int_0^\infty e^{-\frac{E}{k_B T}}\, \Omega(N,V,E)\, dE. \tag{2.9}$$

Unlike the relationships obtained via discrete formulation of partition function where the energy states in the exponential term were the energy levels of individual particles, here E is the total energy for the system of particles. Consequently, the discrete formulation is mostly encountered in quantum-chemistry texts and software tools that treat energy states of single particles, while the continuous formulation is the method of choice for studying a collection of particles, e.g. in molecular dynamics or Monte-Carlo simulation.

For its more common employment in theoretical chemistry and more frequent applications throughout this thesis, in the following we introduce the evaluation of partition functions based on the discrete formulation. The continuous formulation of statistical thermodynamics will be further discussed in section 2.4.

[2] A system with constant number of particles, volume, and energy

[3] A system with constant number of particles, volume, and temperature

2.2 Evaluation of partition functions

As implied from the physical interpretation of the partition function written as $Q = \sum_i g_i e^{-\frac{\varepsilon_i}{kT}}$, its determination requires to find all possible energy states ε_i observable for a particle and their degeneracy and computing the summation. Additionally, we can also include the degeneracy in the summand to write the partition function as:

$$Q = \sum_i e^{-\frac{\varepsilon_i}{k_B T}}, \tag{2.10}$$

which is more common in statistical thermodynamics textbooks. Here, members of a degenerate state get different values of the index i. Nevertheless, the final results are the same, the only difference is in the indexing of the energy states.

To be able to obtain analytical relationships for the evaluation of partition functions, we can decompose each energy state into a combination of different contributions, from kinetic energy, which includes translational and rotational motions, and potential energy, which includes vibrational and electronic energies. This can be written as:

$$\begin{aligned} Q &= \sum_i e^{-\frac{\varepsilon_{trans.,i}+\varepsilon_{rot.,i}+\varepsilon_{vib.,i}+\varepsilon_{elect.,i}}{k_B T}} \\ &= \sum_i e^{-\frac{\varepsilon_{trans.,i}}{k_B T}} \sum_i e^{-\frac{\varepsilon_{rot.,i}}{k_B T}} \sum_i e^{-\frac{\varepsilon_{vib.,i}}{k_B T}} \sum_i e^{-\frac{\varepsilon_{elect.,i}}{k_B T}} \\ &= Q_{trans.} Q_{rot.} Q_{vib.} Q_{elect}\,. \end{aligned} \tag{2.11}$$

To evaluate each one of the partition function components for a given molecule, an extensively employed and successful approach is to consider the molecules as rigid rotors in which the vibrations follow the harmonic oscillator model. This is known as the Rigid Rotor Harmonic Oscillator (RRHO) approximation. It allows us to employ quantum-mechanically defined relationships describing various energy components for RRHO and to approximate the sum by integration.

For the translational energy, possible energy states in one-dimensional motion of a particle with mass M in a box with l side length, are defined in quantum mechanics as [5]:

$$\varepsilon_{trans.,i} = \frac{i^2 h^2}{8Ml^2}, \tag{2.12}$$

where $i = 1, 2, \ldots$ is the quantum number of discrete bound states. By substituting this relationship in $\sum_i e^{-\frac{\varepsilon_{trans.,i}}{k_B T}}$ and approximating the sum by integration, for three dimensions we obtain [3]:

$$Q_{trans.} = \left(\frac{2\pi M k_B T}{h^2}\right)^{3/2} V, \tag{2.13}$$

where V is the volume of the system.

Similarly, from the quantized rotational energy states of rigid rotors, which by quantum mechanics are defined as [5]:

$$\varepsilon_{rot.,i} = \frac{i(i+1)\hbar^2}{2I}, \tag{2.14}$$

where I is the principal moment of inertia and $i = 0, 1, 2, \ldots$ is a quantum number specifying various rotational energy levels, we can obtain [3]:

$$Q_{rot.} = \frac{\sqrt{\pi}}{s}\left(\frac{8\pi^2 I_A k_B T}{h^2}\right)^{1/2}\left(\frac{8\pi^2 I_B k_B T}{h^2}\right)^{1/2}\left(\frac{8\pi^2 I_C k_B T}{h^2}\right)^{1/2}, \tag{2.15}$$

for non-linear molecules and

$$Q_{rot.} = \frac{8\pi^2 I k_B T}{s h^2}, \tag{2.16}$$

for linear molecules. Here, the pre-factor s is the symmetry number and can play a very important role, especially in isotope-exchange reactions for which isotope exchange results in changing the symmetry of the molecule. For example, for the isotope exchange reaction H_2O+$D_2O \leftrightarrow$ 2HDO with experimentally determined equilibrium constant of 3.75 [6], the isotope exchange results in a dramatic reduction in the product of symmetry numbers from 4 in reactants to 1 in products. For this reaction, by theoretical evaluation of partition functions at DSDPBEP86 (D3)/Def2QZVP level of theory and appropriate treatment of the symmetry number, I could obtain an equilibrium constant equal to 3.8526 which was in excellent agreement with the experimentally determined data. Nevertheless, for the same computations, overlooking the symmetry number results in an equilibrium constant of 0.963, which is far below the actual value and shows the importance of the symmetry number. Noteworthy, although the symmetry number was considered initially in an empirical way and

based on spectroscopy measurement conventions, there are rigorous theoretical justifications for it, based on both quantum mechanics [5] and classical mechanics [7].

For the vibrational partition function, using the quantized energy states of a harmonic oscillator, which in quantum mechanics is described by:

$$\varepsilon_{vib.,i} = \sum_j \hbar\omega_j(n_i + \frac{1}{2}), \tag{2.17}$$

where ω_j is the angular frequency for the jth mode of vibration, the vibrational partition function is obtained as:

$$Q_{vib.} = \prod_i \frac{e^{-\frac{hc\nu_i}{2k_BT}}}{1 - e^{-\frac{hc\nu_i}{k_BT}}}, \tag{2.18}$$

in which the angular frequencies are converted to normal mode vibrational frequencies ν_i, as they are more commonly encountered in spectroscopy. Theoretical evaluation of the normal mode vibrational frequency is known as normal mode analysis and will be discussed in more detail in section 2.3.

For the electronic partition function, the energy gap between the ground and excited states typically is so large that, except for very high temperatures, the Boltzmann factor associated with the excited states can be neglected, compared to that of the ground state. Consequently, in standard thermochemistry computations, only the ground state of the electronic energy is considered, resulting in the electronic partition function written as:

$$Q_{elect} = ge^{-\frac{\varepsilon_{elect.,ground}}{k_BT}}. \tag{2.19}$$

In the above relationship, the pre-factor g, which is the degeneracy of the ground-state energy, is equal to unity for closed-shell particles. However, for other cases, it will be greater than unity and should be taken into account [5].

In the above, I provided a brief introduction to general concepts in statistical thermodynamics and evaluation of partition functions commonly encountered in theoretical chemistry. We employ these obtained partition functions to estimate the enthalpy of combustion reactions in chapter 3 and the equilibrium constant of an isotope exchange reaction in chapter 4. For approximation of partition functions based on the continuous formulation of statistical thermodynamics which is typically achieved via molecular dynamics or Monte-Carlo simula-

tion, different methods such as thermodynamic integration, alchemical perturbation, WHAM, and several other methods are commonly employed. As an excellent introductory textbook for those methods, the interested readers can refer to an excellent introductory textbook written by Chipot and Pohorille [8].

2.3 An introduction to normal-mode analysis

As we saw in section 2.2, evaluation of normal-mode vibrational frequencies is the central step in the calculation of the vibrational partition function. For that purpose, normal-mode analysis based on the harmonic oscillator approximation is commonly the method of choice in computational chemistry, which will be briefly introduced in the following.

If the potential energy for one-dimensional oscillation in the x direction is denoted by $U(x)$, by its Taylor expansion around any arbitrary point x_0 up to second order, we obtain:

$$U(x) \approx U(x_0) + \frac{dU}{dx}(x - x_0) + \frac{1}{2!}\,\frac{d^2U(x)}{dx^2}(x - x_0)^2 . \tag{2.20}$$

Since the potential energy is expanded only up to the second-order, it is called the harmonic potential due to its harmonic shape around the minimum of the potential energy, as depicted in figure 2.2.

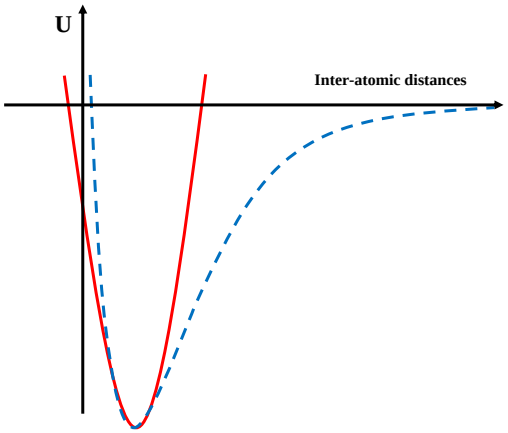


Fig. 2.2 Comparison of harmonic potential (red) and the actual potential (dashed line) in a diatomic molecule. As can be seen, for the distances close to the equilibrium which is located as the minima in the two graphes, the harmonic potential is a good approximation of the real potential.

By taking the point x_0 to be the equilibrium distance, defined as the point at which the potential energy of vibration is minimal ($\frac{dU}{dx} = 0$), and selecting a reference point of potential energy for which $U\left(x_{eq}\right) = 0$, Eq. (2.20) can be rewritten as:

$$U(x) = \frac{1}{2}K(x - x_{eq})^2, \tag{2.21}$$

in which $\frac{d^2U(x)}{dx^2}$ is denoted by K and is called the force or spring constant.

This simply found relationship is the cornerstone of normal mode analysis, as demonstrated in the following. Before going into details, I would like to draw the readers' attention to why the first step in frequency calculation based on the normal mode analysis is always geometry optimization. The reason is that for the optimized geometry, we find a molecular structure for which the first derivative of the potential energy with respect to atomic coordinates is zero, which is indeed the requirement to arrive at Eq. (2.21) from the initially derived Taylor expansion. Additionally, the introduced derivation implies that when we are dealing with a collection of molecules and are interested in evaluating the normal modes for only one of them, we do not necessarily need to optimize the geometry for all molecules and can limit the geometry optimization to the individual molecule of interest only. This was the premise of developing the *partial* normal mode analysis presented in chapter 4, which was employed as an effective strategy to significantly reduce the computational cost of normal mode analysis.

To evaluate the motion of atoms by the harmonic potential described by Eq. (2.21), we can employ Newton's second law of motion written as $F = m\frac{d^2(x-x_{eq})}{dt^2}$, and the connection between F and $U(x)$ defined in classical mechanics as $F = -\frac{dU(x)}{dx} = -K\left(x - x_{eq}\right)$, which results:

$$-K\left(x - x_{eq}\right) = m\frac{d^2(x - x_{eq})}{dt^2}. \tag{2.22}$$

To make the mathematical formulation more abbreviate, we can define new variables, namely mass-weighted coordinates written as $q = \sqrt{m}\,(x - x_{eq})$ and a mass-weighted force constant defined as $K^{'} = \frac{K}{m}$, and substitute them in Eq. (2.22) to obtain:

$$-K^{'}q = \frac{d^2q}{dt^2}. \tag{2.23}$$

This obtained differential equation has an analytical solution of the form:

$$q = a\cos\left(\sqrt{K'}\,t\right), \tag{2.24}$$

where a is a constant. This equation clearly shows a periodic oscillation for the particle as a function of time with a period of $\frac{\sqrt{K'}}{2\pi}$, which can be regarded as the frequency with dimension $\frac{1}{time}$. We can generalize this methodology to multi-atomic molecules. For that purpose, instead of a force constant for a single coordinate, we need a matrix of force constants for which the ij elements are $\frac{\partial^2 U}{\partial q_i \partial q_j}$ and i and j span all 3N coordinates in a molecule containing N atoms, which is the so-called Hessian matrix. The Hessian matrix elements can be analytically or numerically determined by quantum mechanical computations. After that, by diagonalizing the Hessian matrix in the mass-weighted coordinates, we obtain a set of eigenvalues, the square roots of which are the normal-mode frequencies as implied from Eq. (2.24), and eigenvectors for the corresponding modes of vibration. The diagonalization employed here, in fact implements a coordinate transformation so that the off-diagonal elements of the Hessian matrix become zero. By this transformation, we can remarkably reduce the computational complexity without losing accuracy.

2.4 Partition function for ensemble of particles: an introduction to molecular dynamics simulation

As it was discussed in section 2.1, the thermodynamic quantities are in fact statistical averages of those quantities in all possible microstates, weighted by the Boltzmann factor. For single molecules in vacuo, we have shown how the rigid rotor harmonic oscillator (RRHO) approximation can be employed to approximate the ensemble average of microstates attributed to different components of energy. Nevertheless, calculations based on the RRHO approximation are typically limited to small molecules that can be approximately treated as rigid bodies. For large molecules, rotations around backbone dihedrals can yield diverse configurations with quite different principal moments of inertia, thus clearly violating the rigid rotor approximation. Additionally, the original implementations of static computations based on the RRHO approximation are commonly limited to in-vacuo states of single particles. Although taking into account solvent effects became possible in the past few decades via the continuum solvation approach introduced in section 2.6, nevertheless, this approach also is just a model with considerable simplifications and parameterizations, as discussed in detail in chapter 8.

Obviously, for any system of interest, the exact partition function can be calculated by generating all possible configurations, calculating the resulting energy states, converting them to the corresponding probabilities based on the Maxwell-Boltzmann statistics, and using those obtained probabilities to compute the statistical average of any observable. For a collection of particles, this can be mathematically formulated as the integral introduced in Eq. (2.8). By this integration, we in fact try all possible combinations of momentum and spatial coordinates for all particles in the system and calculate the resulting energy, which spans all possible energy states of the system, as reflected in the total Hamiltonian of the system. The calculated Hamiltonian can then be used to approximate the probability of each microstate.

Although with studying all possible energy states we can exactly evaluate the partition function and its attributed thermodynamic quantities, nevertheless, this exact solution is only feasible for small systems consisting of very few particles. For larger systems, the total number of possible microstates soon approaches astronomically large numbers, which makes exact evaluation of all microstates in practice impossible.

To overcome this limitation, a concept can be exploited that is known as importance sampling in statistical thermodynamics. To clarify the concept of importance sampling, we can consider a simple example of studying one H_2 molecule in a large box. For this example, a large portion of possible configurations is those for which the hydrogen atoms are so far away from each other that there is no bond between them. However, for all but very high temperatures, we know that such configurations almost never occur. This is also consistent with our theoretical formulation because the Boltzmann factor for such configurations approaches zero, which causes those configurations to have no contribution in the ensemble average of required quantities.

The general idea behind importance sampling is that instead of studying all possible configurations, which in most cases have very negligible probabilities to occur, we only generate and study configurations that are likely to be visited in practice. For that purpose, a number of efficient computer simulations have been developed and can be used to generate and study those important configurations. In this section, we briefly introduce one of the most widely used computer simulation algorithms of this kind, which is molecular dynamics simulation. A quantum-mechanical variant of molecular dynamics simulation, namely the path integral molecular dynamics simulation, will be introduced in the next section. Yet another extensively used computer simulation algorithm for importance sampling is the Monte-Carlo simulation which will not be discussed here, as in will not be employed it this thesis. Readers who are interested to know more about this method are referred to the book

Understanding Molecular Simulation: From Algorithms to Applications [9], as one of the excellent introductory texts to the Monte-Carlo simulation technique.

As was mentioned above, molecular dynamics simulation is one of the most widely employed importance sampling algorithms. The general idea behind molecular dynamics simulation is to set up a simulation box as a small (or periodic) representation of the system of interest. Then, based on the position of the particles in each step, the net force exerted on each particle, which depends on the interatomic distances between them, can be evaluated, either via quantum-mechanical computations or algebraic relationships that are referred to as force fields. The calculated forces in addition to the given speed and position of all particles in each step are then used to estimate the speeds and positions of the particles in the next time step, based on Newtonian mechanics. These computations are repeated for several steps to generate a trajectory for the temporal evolution of the system. This trajectory, which consists of samples that have important contributions in the ensemble average of quantities, is used to compute the partition function and various quantities attributed to it.

To prevent this section from becoming too lengthy, we avoid providing details on the algorithms of molecular dynamics simulations and refer the interested readers to several available textbooks, such as Understanding Molecular Simulation written by Frenkel and Smit [9] and Computer simulation of liquids written by Allen and Tildesley [10].

2.5 PIMD Simulation: an introduction

In the previous section, we briefly introduced the general idea behind molecular dynamics simulations. Although very robust in studying many aspects of classical-mechanical systems, molecular dynamics simulation suffers from the drawback that it relies on Newtonian mechanics and hence ignores quantum effects by construction. This can be problematic when studying systems for which quantum effects are significant, such as those dealing with low temperatures or light particles like hydrogen transfer reactions as well as isotope exchange reactions. For such cases, taking into account quantum effects in the framework of molecular dynamics simulations has become possible through a method which is called path integral molecular dynamics. In the following, I present a quick and straightforward introduction to the basics of path integral molecular dynamics simulation as one of the tools employed in studying isotope effects in chapter 4. For further details on this method, the interested readers can refer to excellent books written by Feynman [11] and Tuckerman [4].

Deriving the path integral molecular dynamics starts from the time-dependent Schrödinger equation written as:

$$i\hbar\frac{\partial}{\partial t}\,|\Psi(t)\rangle = \hat{H}\,|\Psi(t)\rangle\,, \tag{2.25}$$

where $\hat{H}$ is the Hamiltonian operator and $\Psi(t)$ is the time-dependent wave function of the system, and its formal solution given by:

$$|\Psi(t)\rangle = e^{-i\hat{H}(t-t_0)/\hbar}\,|\psi(t_0)\rangle\,. \tag{2.26}$$

In position representation with coordinate q, the solution of the time-dependent Schrödinger equation can be stated as:

$$\begin{aligned}
\Psi(q,t) = \langle q|\psi(t)\rangle &= \langle q|e^{-i\hat{H}(t-t_0)/\hbar}|\psi(t_0)\rangle \\
&= \int dq_0\,\langle q|e^{-i\hat{H}(t-t_0)/\hbar}|q_0\rangle\,\langle q_0|\psi(t_0)\rangle \\
&= \int dq_0\,\langle q|e^{-i\hat{H}(t-t_0)/\hbar}|q_0\rangle\,\psi(q_0,t_0) \\
&= \int dq_0\,K(q,q_0,t)\psi(q_0,t_0)\,,
\end{aligned} \tag{2.27}$$

where the resolution of the identity operator $\int dq_0\,|q_0\rangle\langle q_0| = 1$ was used in lines 2 and 3 of the above derivation. The term $K(q,q_0,t) = \langle q|e^{-i\hat{H}(t-t_0)/\hbar}|q_0\rangle$ is a kernel which plays the central role in the path integral formalism. For a free particle in one dimension with the Hamiltonian written as $\hat{H} = \frac{\hat{P}^2}{2m}$, where $\hat{P} = -i\hbar\frac{\partial}{\partial q}$ is the quantum-mechanical momentum operator, for $t_0 = 0$ as an arbitrary choice and using the inner product of momentum and position operators defined as $\langle q|p\rangle = \frac{e^{\frac{ipq}{\hbar}}}{\sqrt{h}}$, this kernel can be written as:

$$\begin{aligned}
K^{\text{free}}(q,q_0,t) &= \langle q|e^{\frac{-it}{\hbar}\frac{\hat{P}^2}{2m}}|q_0\rangle \\
&= \int_{-\infty}^{\infty} dp\;\langle q|e^{\frac{-it}{\hbar}\frac{\hat{P}^2}{2m}}|p\rangle\langle p|q_0\rangle \\
&= \int_{-\infty}^{\infty} dp\,\langle q|p\rangle\,e^{\frac{-it}{\hbar}\frac{p^2}{2m}}\;\langle p|q_0\rangle \\
&= \frac{1}{h}\int_{-\infty}^{\infty} dp\,e^{\frac{ipq}{\hbar}}\,e^{\frac{-it}{\hbar}\frac{p^2}{2m}}\,e^{\frac{-ipq_0}{\hbar}} \\
&= \frac{1}{h}\int_{-\infty}^{\infty} dp\,e^{\frac{-i}{\hbar}p(q_0-q)}\,e^{\frac{-it}{\hbar}\frac{p^2}{2m}}\,,
\end{aligned} \tag{2.28}$$

which via completing the square and using the Fresnel integral equation can be analytically solved to yield:

$$K^{\text{free}}(q, q_0, t) = \frac{e^{\frac{im}{2\hbar t}\,(q_0-q)^2}}{\sqrt{ith/m}}\,. \tag{2.29}$$

The obtained kernel allows finding the solution of a more general case, which is a particle under a time-independent potential $V(q)$, for which the kernel of interest is defined as:

$$K(q, q_0, t) = \langle q|e^{\frac{-it}{\hbar}\left(\frac{\hat{p}^2}{2m}+\mathrm{V}(\hat{q})\right)}|q_0\rangle\,. \tag{2.30}$$

To that end, using the Trotter product formula which states:

$$e^{\hat{A}+\hat{B}} = \lim_{\mathrm{P}\to\infty}\left(e^{\hat{A}/\mathrm{P}}e^{\hat{B}/\mathrm{P}}\right)^{\mathrm{P}}\,, \tag{2.31}$$

we can write:

$$K(q, q_0, t) = \lim_{\mathrm{P}\to\infty}\langle q|\left(e^{\frac{-it}{\mathrm{P}\hbar}\frac{\hat{p}^2}{2m}}\,e^{\frac{-it}{\mathrm{P}\hbar}\,\mathrm{V}(\hat{q})}\right)^{\mathrm{P}}|q_0\rangle\,, \tag{2.32}$$

which by inserting $\mathrm{P}-1$ resolution of the identity operators $\int dq_i\,|q_i\rangle\langle q_i| = 1$ can be rewritten as:

$$\begin{aligned} K(q, q_0, t) &= \lim_{\mathrm{P}\to\infty}\int dq_1, \ldots, dq_{\mathrm{P}-1}\prod_{i=0}^{\mathrm{P}-1}\langle q_{i+1}|\,e^{\frac{-it}{\mathrm{P}\hbar}\frac{\hat{p}^2}{2m}}e^{\frac{-it}{\mathrm{P}\hbar}\,\mathrm{V}(\hat{q}_i)}|q_i\rangle \\ &= \lim_{\mathrm{P}\to\infty}\int dq_1, \ldots, dq_{\mathrm{P}-1}\prod_{i=0}^{\mathrm{P}-1}\langle q_{i+1}|\,e^{\frac{-it}{\mathrm{P}\hbar}\frac{\hat{p}^2}{2m}}|q_i\rangle e^{\frac{-it}{\mathrm{P}\hbar}\,\mathrm{V}(q_i)}\,. \end{aligned} \tag{2.33}$$

Here, the term $\langle q_{i+1}|e^{\frac{-it}{\mathrm{P}\hbar}\frac{\hat{p}^2}{2m}}|q_i\rangle$ is equivalent to $K^{\text{free}}(q, q_0, t)$ defined in Eq. (2.28) and by inserting its analytical solution given by Eq. (2.29) we finally obtain:

$$K(q, q_0, t) = \lim_{\substack{\mathrm{P}\to\infty\\ \delta t\to 0}}\left(\frac{m}{hi\delta t}\right)^{\mathrm{P}/2}\int dq_1, \ldots, dq_{\mathrm{P}-1}\; e^{\frac{i}{\hbar}\left[\sum_{i=0}^{\mathrm{P}-1}\frac{m\delta t}{2}\left(\frac{q_i-q_{i+1}}{\delta t}\right)^2-\mathrm{V}(q_i)\delta t\right]}, \tag{2.34}$$

where $\frac{t}{\mathrm{P}}$ is replaced by δt. By inserting this found kernel in Eq. (2.27) we get:

$$\Psi(q,t) = \int dq_0 \lim_{\substack{\mathrm{P}\to\infty \\ \delta t\to 0}} \left(\frac{m}{hi\delta t}\right)^{\mathrm{P}/2} \int dq_1,\ldots,dq_{\mathrm{P}-1}\; e^{\frac{i}{\hbar}\left[\sum_{i=0}^{\mathrm{P}-1} \frac{m\delta t}{2}\left(\frac{q_i-q_{i+1}}{\delta t}\right)^2 - \mathrm{V}(q_i)\delta t\right]}\; \Psi(q_0,t_0)\,, \tag{2.35}$$

which can be interpreted as the summation of all possible paths a particle located in q_0 at time t_0 can follow to get to the point q_t at time t, schematically illustrated in figure 2.3. Furthermore, as can be implied from the found relationship, the probabilities of following each one of the possible paths are taken into account via the weights given by the exponential term. Due to the summation of all possible paths, this formalism is called *path integral*.

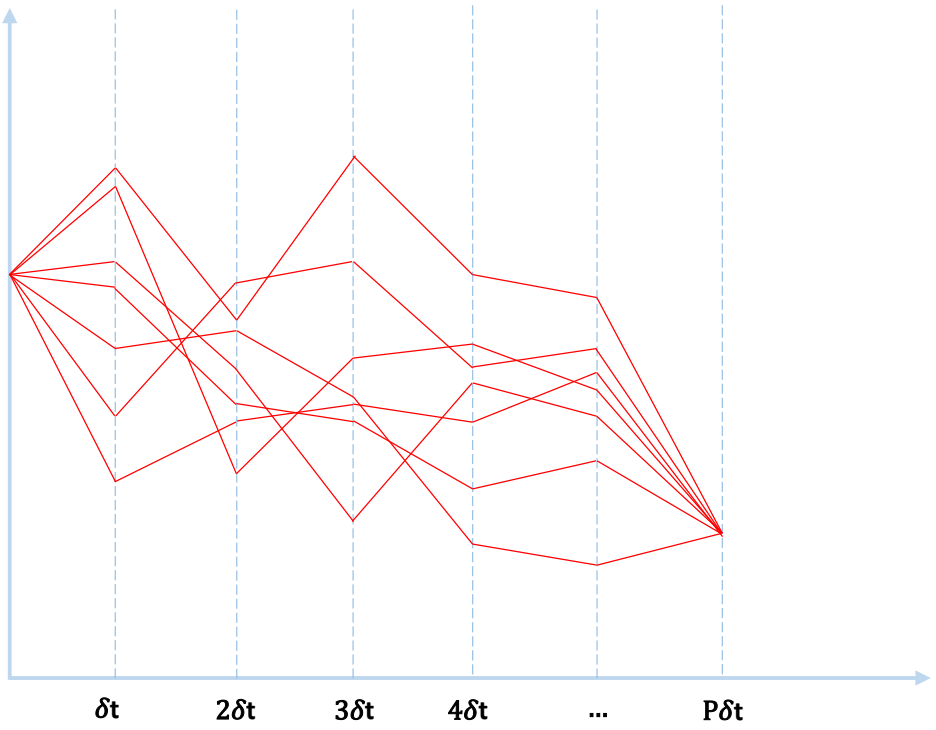


Fig. 2.3 Illustration of the different possible paths followed by a particle to move between two points.

We can use the path integral formalism to evaluate the quantum-mechanical partition function, which in analogy to its classical counterpart introduced in section 2.1 is defined as

$Q = \sum_i \langle\psi_i|e^{-\beta H}|\psi_i\rangle$, where $\beta = \frac{1}{k_B T}$ which in position representation results in the configuration partition function (Z) written as:

$$Z = \int d\mathrm{q}\, \langle q|e^{-\beta H}|q\rangle \,. \tag{2.36}$$

Employing the Trotter factorization similar to the way it was applied to derive the path integral formalism results:

$$Z = \lim_{\mathrm{P}\to\infty} \int dq_1, \ldots, dq_{\mathrm{P}-1} \prod_{i=1}^{\mathrm{P}} \langle q_{i+1}|e^{-\beta_P \frac{\hat{p}^2}{2m}}\, e^{-\beta_P\, \mathrm{V}(\hat{q}_i)}|q_i\rangle, \tag{2.37}$$

where $\beta_P = \frac{\beta}{\mathrm{P}}$. Using the same manipulations employed to obtain Eq. (2.29), we can show that:

$$\langle q_{i+1}|e^{-\beta_P \frac{\hat{p}^2}{2m}}\, e^{-\beta_P\, \mathrm{V}(\hat{q}_i)}|q_i\rangle = \frac{1}{2\pi\hbar}\sqrt{\frac{2\pi m}{\beta_P}}\, e^{-\beta_P\left[\frac{m}{2\hbar^2\beta_P{}^2}\,(q_{i+1}-q_i)^2 +\, \mathrm{V}(q_i)\right]}, \tag{2.38}$$

which by defining $\omega_p = \frac{1}{\beta_P \hbar}$ and substituting in Eq. (2.37) finally results in:

$$Z = \lim_{\mathrm{P}\to\infty} \left(\frac{1}{2\pi\hbar}\right)^{\mathrm{P}} \left(\frac{2\pi m}{\beta_P}\right)^{\mathrm{P}/2} \int dq_1, \ldots, dq_{\mathrm{P}-1} e^{-\beta_P \sum_{i=1}^{\mathrm{P}}\left[\frac{m{\omega_p}^2}{2}\,(q_{i+1}-q_i)^2 +\, \mathrm{V}(q_i)\right]}. \tag{2.39}$$

The found configurational partition function described by Eq. (2.39) plays the central role in path integral molecular dynamics simulation. It can be interpreted as a ring-polymer of classical particles where the time evolution of each particle follows classical molecular dynamics. Nevertheless, the ring-polymer Hamiltonian that appears in the exponential term results in the sampled phase space including quantum effects, as we saw in the derivation of the path integral formalism.

In analogy to classical mechanics and the harmonic oscillation introduced in section 2.3, each one of the P replicas can be considered as some beads that construct the path integral ring-polymer and are connected by a harmonic spring with force constant $m{\omega_p}^2$. Accordingly, ω_p can also be interpreted as the ring-polymer frequency. A schematic representation of the path integral ring-polymer for a water molecule with 3 beads is depicted in figure 2.4.

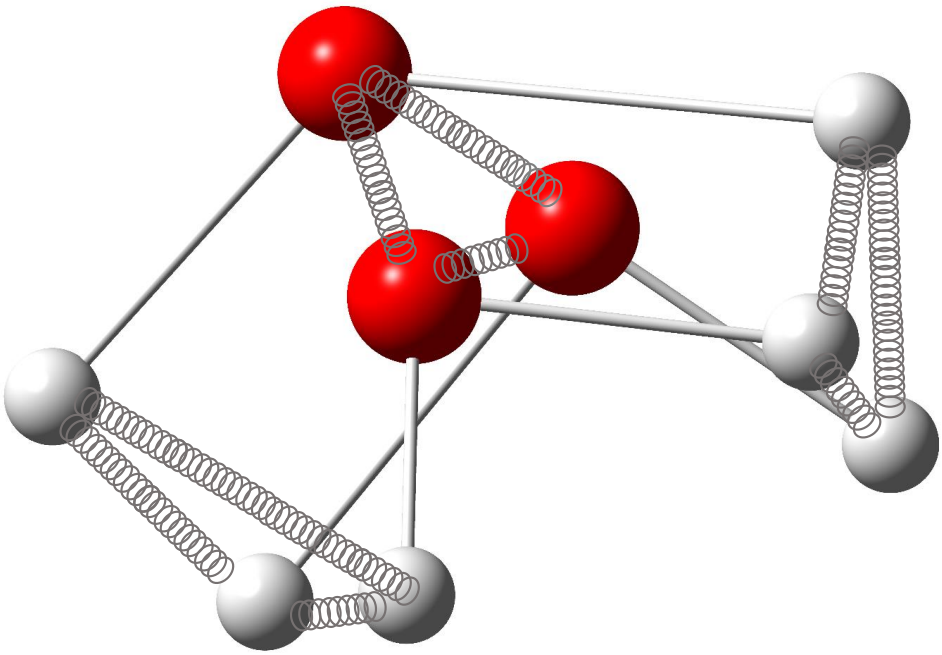


Fig. 2.4 Three-bead ring-polymer illustration of a water molecule. This selected number of beads is chosen only for the clarity of the image and is not related to the number of atoms in the water molecule.

As implied from this force constant, for very light masses or low temperatures at which $\omega_p = \frac{1}{\beta_P \hbar}$ is a very small number, the spring constant becomes a small number and results in expansion of the ring polymer, intuitively analogous to an increased Heisenberg position uncertainty. On the other hand, for heavy particles or high temperatures, this spring constant becomes a large number, and the stronger spring results in a collapse of all the beads to a single point, analogous to a single classical-mechanical particle. This shows that quantum effects are included mainly via the stiffness of the spring connecting the ring-polymer beads, as can be inferred from the figure 2.4.

In the present thesis, we employ path integral molecular dynamics simulations to study isotope effects for equilibrium between boric acid and borate in pure water and seawater, presented in chapter 4.

2.6 Studying solvent effects by implicit solvent approaches

Taking into account the solvent effect is one of the vibrant areas in theoretical and computational chemistry and is widely required, due to a large number of the important physical-chemistry phenomena happening in solution. For that purpose, two main approaches, namely

implicit and explicit solvent approaches, can be employed. The features and pros and cons of these two approaches are discussed in chapter 8. In the current section, I briefly introduce some important theoretical backgrounds to the implicit solvent approach which are mainly taken from the textbooks Essentials of Computational Chemistry [12] and Introduction to Computational Chemistry [13]. For a more complete overview of the topic, the interested readers can refer to those textbooks.

Based on the implicit solvation approach, instead of direct evaluation of the charge distribution of solutes with that of the solvent, a continuous reaction field is defined that represents the statistical average of electric fields due to various configurations of the solvent. The solvent effects then are taken into account via studying the interactions between the charge distribution in the solute and the reaction field of the solvent.

If a molecule is subjected to an electric field definable as the gradient of the electrostatic potential (ϕ), from basic electrostatic physics the total energy resulting from the interaction of the charge distribution in that molecule $\rho(r)$ and the electric field is obtained as:

$$W = -\frac{1}{2}\int \rho(r)\,\phi(r)\,dr\,, \tag{2.40}$$

which is equivalent to the work required for bringing that molecule from vacuum with no electric field to the point in which the electric field is present. Based on this definition, the polarization energy in the implicit solvation is also defined as the difference of this computed energy between the gas and solution states. From this viewpoint, various implicit solvation models are in fact different in how to define the charge distribution, the electric field, and their interactions.

One of the classical models for studying the charge distribution and the reaction field is the Poisson equation, which is defined as:

$$\nabla.(\varepsilon(r).\nabla\phi(r)) = -4\pi\rho(r)\,, \tag{2.41}$$

in which $\varepsilon(r)$ is the dielectric constant of the solvent. The Poisson equation holds for solutions with negligible ionic strengths. For a more general case applicable also to electrolytes, an extension of the Poisson equation, which is known as the Poisson-Boltzmann equation, can be employed, which is defined as:

$$\nabla(\varepsilon(r).\nabla\phi(r)) - \varepsilon(r)\lambda(r)\kappa^2\frac{k_BT}{q}\sinh\left[\frac{q\phi(r)}{k_BT}\right] = -4\pi\rho(r), \tag{2.42}$$

where $\lambda(r)$ is a parameter equal to one, except for the regions inaccessible by the solute, in which it has a value of zero, q represents the ionic charges in the electrolyte, and κ^2 is the Debye– Hückel parameter, which is defined as:

$$\kappa^2 = \frac{8\pi q^2}{\varepsilon k_BT}. \tag{2.43}$$

To further simplify the Poisson-Boltzmann equation, for low ionic strength solutions we can approximate the hyperbolic sine with a truncated power expansion to obtain:

$$\nabla(\varepsilon(r).\nabla\phi(r)) - \varepsilon(r)\lambda(r)\kappa^2\,\phi(r) = -4\pi\rho(r), \tag{2.44}$$

which is called the linearized Poisson-Boltzmann equation. The other classical model, applicable for evaluation of the polarization energy in monoatomic ions, is the Born model, which is based on considering the uniform charge density on the atomic surface, written as:

$$\rho = \frac{q}{4\pi a^2}, \tag{2.45}$$

where q is the ionic charge and a is the atomic radius. Based on the Gauss law of electric fields, the electrostatic potential at the distance r from this atom is obtained via:

$$\phi(r) = -\frac{q}{\varepsilon r}, \tag{2.46}$$

which by substitution in Eq. (2.40) results in:

$$W = -\frac{1}{2}\int\frac{q}{4\pi a^2}\left(-\frac{q}{\varepsilon a}\right)ds = \frac{q^2}{2\varepsilon a}, \tag{2.47}$$

and the polarization energy equal to:

$$G_p = -\frac{1}{2}(1-\frac{1}{\varepsilon})\frac{q^2}{a}\,, \tag{2.48}$$

known as the Born model of polarization.

If we do the same calculations for a perfectly dipolar charge distribution with dipole moment μ instead of the employed uniform charge distribution, we obtain:

$$G_p = -\frac{1}{2}(\frac{2(\varepsilon-1)}{2\varepsilon+1})\frac{\mu^2}{a^3}\,, \tag{2.49}$$

which is called the Kirkwood-Onsager polarization energy. To estimate the Kirkwood-Onsager polarization energy by quantum mechanical computations, Eq. (2.49) can be inserted into the Fock equation, resulting in:

$$\left\{F_i - (\frac{2(\varepsilon-1)}{2\varepsilon+1})\frac{\langle\psi|\mu|\psi\rangle^2}{a^3}\right\}\psi_i = e_i\psi_i\,, \tag{2.50}$$

in which F_i is the Fock operator for molecular orbital i in the gas phase. This resulting non-linear Fock equation can be solved iteratively, which is known as self-consistent reaction field calculations. Further details on self-consistent reaction field calculations are discussed in chapter 8.

2.7 Machine learning

The methods introduced in the previous sections mainly are mathematical approaches that can be employed for the theoretical study of the physics behind a problem. However, by looking into nature, we immediately notice that many living entities are able to learn and solve problems that in most cases are far more complicated than the ones which are the aim of the present study, interestingly without any explicit employment of mathematics. For example, a human child in the earliest years of its life, even before gaining the ability to speak, can learn to distinguish many fruits from each other. If we want to write a computer algorithm for that purpose, it would probably require several thousand lines of code with a lot of mathematics to define and calculate the geometry of different fruits, parameters like ratios of cross-section to length, colors, and so on. The process in which living entities learn to distinguish between features is learning via examples. For the fruit example, a human

baby acquires this ability by being provided several examples of different fruits along with their names. This learning analogy stimulated designing algorithms for similar learning by computers, i.e. learning by providing examples, and is referred to as machine learning.

Through several decades of advancements, machine learning has now become a highly powerful and reliable tool to study the most complicated problems in many scientific fields. Examples of applications of machine learning in chemistry and life science are provided in chapter 9. In this section, I briefly introduce some basics of machine learning required to study a wide range of problems of interest in chemistry.

To develop a machine-learning model, as the first step we are required to translate the structure of molecules to a set of numerical data which are called molecular representations or descriptors. One of the earliest examples of employing such descriptors is the group contribution method, or equivalently the structure-property relationship method, which was proposed in the middle of the 20$^{\text{th}}$ century [14]. According to the group contribution method, a molecule is quantitatively described by a set of numbers that specify the number of each previously defined functional group. For example, in chapter 6, we define 42 functional groups and each molecule is characterized by a vector with 42 elements. Each element records the number of presences of a specific functional group with the same ID in the studied molecule. Nevertheless, for more challenging problems, more versatile molecular representations are commonly required. Some of the most widely applied molecular representations are briefly introduced in the following.

2.7.1 Atom-centered symmetry functions

Atom-centered symmetry functions were proposed by Behler and Parrinello in 2007 for numerical characterization of chemical environments and as inputs to machine-learning models they developed for evaluating the potential energy [15]. Atom-centered symmetry functions can be defined either as radial functions given as:

$$G_i^{atom,\ radial} = \sum_{j=1}^{N_{atoms}} e^{-\eta(r_{ij}-r_s)^2} f_c\left(r_{ij}\right), \tag{2.51}$$

or angular functions:

$$G_i^{atom,\ angular} = 2^{1-\zeta} \sum_{j,k\neq i}^{all} (1+\lambda\cos\theta_{ijk})^{\zeta} e^{-\eta\left({r_{ij}}^2+{r_{ik}}^2+{r_{jk}}^2\right)} f_c\left(r_{ij}\right) f_c(r_{ik}) f_c\left(r_{jk}\right), \tag{2.52}$$

where $f_c\left(r_{ij}\right)$ is a cutoff function defined as:

$$f_c(r_{i,j}) = \begin{cases} 0.5\left[\cos\left(\frac{\pi r_{i,j}}{r_c}\right)+1\right], & \text{for } r_{i,j} \leq r_c \\ 0.0, & \text{elsewhere} \end{cases} \tag{2.53}$$

and r_{ij} is the distance between atoms i and j. For the radial functions, the parameter r_c is the cutoff distance, and the parameters r_s and η shift the Gaussian position and adjust the Gaussian width, respectively. For the angular function, the parameter ζ controls the angular resolution and the parameter λ specifies the location of extrema of the cosine function.

2.7.2 Bispectrum of the neighbor density

Bispectrum of the neighbor density representations were proposed in 2010 by Bartók et al. and represent the chemical space by a series of spherical harmonics [16]. For that purpose, the chemical space is first characterized via neighbor densities $\rho_i(r)$ for each atom i:

$$\rho_i(r) = \delta(r) + \sum_{r_{ij}<r_c} f_c\left(r_{ij}\right) w_j \delta\left(r - r_{ij}\right), \tag{2.54}$$

where δ is the Kronecker delta function, $f_c\left(r_{ij}\right)$ is the same cutoff function introduced in the previous section, and w_j is an element-specific parameter to allow distinguishing between different elements. The calculated neighboring densities are then projected onto a sphere resulting in the projected density written as:

$$\rho(\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\theta,\phi). \tag{2.55}$$

This projected density is then extended to another four-dimensional density function including an additional angle θ_0 that can be expanded by hyper-spherical harmonics $U^j_{m,m'}$ to result in:

$$\rho = \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} c^j_{m,m'} U^j_{m,m'}. \tag{2.56}$$

The coefficients ${c_{m,m'}}^j$ are finally employed to construct the bispectrum matrix with $B_{j_1,j_2,j}$ elements given by:

$$B_{j_1,j_2,j} = \sum_{m_1,m'_1=-j_1}^{j_1} \sum_{m_2,m'_2=-j_2}^{j_2} \sum_{m,m'=-j}^{j} c^{j}_{m,m'} c^{jm}_{m_1 j_1 m_2 j_2} c^{jm'}_{m'_1 j_1 m'_2 j_2} c^{j_1}_{m_1 m'_i} c^{j_2}_{m_2 m'_2}, \tag{2.57}$$

where $c^{jm}_{m_1 j_1 m_2 j_2}$ and $c^{jm'}_{m'_1 j_1 m'_2 j_2}$ are Clebsch-Gordan coefficients.

2.7.3 Smooth overlap of atomic orbitals

The Smooth overlap of atomic orbitals is essentially an extension of the Bispectrum of the neighbor density proposed by the same authors, where the delta function is replaced by Gaussian functions centered on each atom, which allowed redefining the neighbor densities as [17]:

$$\rho(r) = \sum_{i=1}^{N} e^{(-\alpha|r-r_i|^2)}. \tag{2.58}$$

The Smooth overlap of atomic orbitals kernel, which is the overlap of two neighbor densities, is defined as:

$$k(\rho,\rho') = \int dR \left[\int \rho(r)\rho'(rR)dR \right]^n. \tag{2.59}$$

where the exponent n is commonly chosen as 2.

2.7.4 Coulomb matrix

The Coulomb matrix representation was proposed in 2012 by Rupp et al. and is defined as the eigenvalues of a matrix with elements given by [18]:

$$\alpha_{i,j} = \begin{cases} 0.5\ (Z_i)^{2.4}, & \text{for } i = j \\ \frac{Z_i\ Z_j}{r_{ij}}, & \text{for } i \neq j \end{cases}, \tag{2.60}$$

where Z_i is the nuclear charge of atom i and r_{ij} is the inter-atomic distance between atoms i and j. As can be implied from this formulation, the number of columns and rows of the Coulomb matrix is equal to the number of atoms in the system. Therefore, for different

reference molecules, different Coulomb matrices with different sizes are computed, which results in difficulties in their employment by the machine-learning process. To overcome this limitation, the developers of this representation suggested considering a constant-size Coulomb matrix with rows and columns appropriate for the largest system under study. For the smaller systems, the extra lines and columns are then filled with zeroes.

2.7.5 Machine-learning algorithms

Once the required representations are computed for the molecules of interest, the machine-learning algorithms are required to learn the dependency between those representations and the target quantities. The learning process in machine learning can be divided into two general categories, namely unsupervised and supervised learning. Unsupervised learning is mainly employed for clustering applications, in which given samples are divided into arbitrary categories that are not known a priori, based on the similarities and dissimilarities between them.

On the other hand, supervised learning, which is encountered in applications such as classification and function approximation as the most widely-used features of machine learning in molecular sciences, requires providing several examples of the problem of interest. Those known reference data are used to fine-tune the model specifications in a way to be able to correctly learn and reproduce the features of interest.

A large number of different machine-learning algorithms are currently developed and can be used for different applications such as artificial neural networks, support vector machines, kernel methods, and so on. In the present thesis, we will mainly employ artificial neural networks as one of the most powerful machine-learning algorithms for function approximation. Further details on artificial neural networks and practical guidelines for developing rigorous neural network models are provided in chapter 6.

References:

1. Smith, J. M., Introduction to chemical engineering thermodynamics. ACS Publications: 1950.

2. Glicksman, M. E., Principles of solidification: an introduction to modern casting and crystal growth concepts. Springer Science & Business Media: 2010.

3. Mayer, J.; Goeppert, M., Mayer, Statistical Mechanics. John Wiley & Sons, New York 1940.

4. Tuckerman, M., Statistical mechanics: theory and molecular simulation. Oxford university press: 2010.

5. Atkins, P.; De Paula, J., Elements of physical chemistry. Oxford University Press, USA: 2013.

6. Simonson, J. M., The enthalpy of the isotope-exchange reaction: H2O+ D2O= 2HDO at temperatures to 673 K and at pressures to 40 MPa. The Journal of Chemical Thermodynamics 1990, 22 (8), 739-749.

7. Gilson, M. K.; Irikura, K. K., Symmetry numbers for rigid, flexible, and fluxional molecules: theory and applications. The Journal of Physical Chemistry B 2010, 114 (49), 16304-16317.

8. Chipot, C.; Pohorille, A., Free energy calculations. Springer: 2007.

9. Frenkel, D.; Smit, B., Understanding molecular simulation: from algorithms to applications. Elsevier: 2001; Vol. 1.

10. Allen, M. P.; Tildesley, D. J., Computer simulation of liquids. Oxford university press: 2017.

11. Feynman, R. P.; Hibbs, A. R.; Styer, D. F., Quantum mechanics and path integrals. Courier Corporation: 2010.

12. Cramer, C. J., Essentials of computational chemistry: theories and models. John Wiley & Sons: 2013.

13. Jensen, F., Introduction to computational chemistry. John wiley & sons: 2017.

14. Gamson, W.; Watson, K. In National Petroleum News 36, Tech. Sect, 1944; p 1944.

15. Behler, J.; Parrinello, M., Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical review letters 2007, 98 (14), 146401.

16. Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G., Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical review letters 2010, 104 (13), 136403.

17. Bartók, A. P.; Kondor, R.; Csányi, G., On representing chemical environments. Physical Review B 2013, 87 (18), 184115.

18. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A., Fast and accurate modeling of

molecular atomization energies with machine learning. Physical review letters 2012, 108 (5), 058301.

Chapter 3

High precision evaluation of the combustion enthalpy

Project overview and motivation:

Evaluation of enthalpy of combustion reactions is of high scientific and industrial importance. In this chapter, we study the evaluation of combustion enthalpies via the ab-initio computations and normal mode analysis. In addition to the important applications of the considered case study, one of the main motivations to study those reactions is to benchmark intricacies of the employed methods which are classical tools in the theoretical evaluation of thermochemistry. The other motivation is curiosity due to several reports on significant inaccuracies of ab-initio computations of combustion enthalpies in the literature. In this project, we study different aspects of theoretical computations of enthalpy of reactions and discuss good-practice strategies for more rigorous employment of ab-initio computations for that purpose.

Novelty aspects:

- The provided methodologies for correcting the non-ideality impacts
- The employed level of theory for calculation of thermal components of the partition function which was the highest compared to other studies reported in the literature on the same topic
- Most accurate reported results for theoretical computation of combustion enthalpy

Connection to other chapters:

The knowledge acquired by studying the methods and levels of theory was transferred to the more challenging case of theoretical evaluation of isotope fractionation in solution in chapter 4.

Contributions:

Carrying out all the computations, method development, and writing the manuscript.

Publication status:

This work has been submitted to the journal of scientific reports on 02.09.2021 and is under review there. Its preprint is available online (DOI:10.33774/chemrxiv-2021-fvcph).

High precision evaluation of the combustion enthalpy by ab-intio computations

Amin Alibakhshi[1*]
Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, 24118 Kiel, Germany
Corresponding author: alibakhshi@pctc.uni-kiel.de

Abstract

Accurate evaluation of combustion enthalpy is of high scientific and industrial importance. Although ab-initio computation of heat of reactions is one of the promising and well-established approaches in computational chemistry, reliable and precise computation of heat of combustion reactions by ab-initio methods is surprisingly scarce in the literature. A handful of works carried out for this purpose report significant inconsistencies between the ab-initio evaluated and experimentally determined combustion enthalpies and suggest empirical corrections to improve the accuracy of predicted data. With this background, the main aim of the present study is to investigate the reasons behind those reported inconsistencies and propose guidelines for a high-accuracy estimation of heat of reactions via ab-initio computations. Through the provided guidelines, the most accurate results ever reported, with average absolute deviation, mean unsigned error, and correlation coefficient of 1.556 kJ/mole, 0.072% and 0.99999, respectively, is achieved for theoretically computed combustion enthalpies of 40 studied hydrocarbons.
Keywords*: Combustion enthalpy, Heat of reaction, ab initio, quantum mechanics*

3.1 Introduction

Combustion is the key process in many important applications such as power production, transportation, heating, synthesis, and processing of materials [1]. Despite being an active research area for over a century, fully understanding many aspects of the combustion processes is still a scientific challenge [2]. Unraveling these challenges sometimes requires research under special operational conditions. A well-known example is combustion experiments under microgravity conditions, which is one of the ongoing research activities in the international space station [3]. Alongside the experimental researches, theoretical studies have also substantially contributed to unraveling many of the complexities in combustion science. High-level quantum-mechanical computations have been found to be a promising tool in

studying the kinetics of combustion reactions [4-6], elucidating the pathways of combustion reactions [7,8], or studying combustion thermochemistry [9,10].

Nevertheless, examples of successful atomic-scale computations leading to high-accuracy prediction of combustion enthalpy as one of the most widely required features of combustion reactions are surprisingly scarce in the literature. In few studies carried out so far for this purpose, substantial inconsistencies have always been reported between the theoretically predicted and experimentally determined combustion enthalpies, as discussed in the following.

Whyman et al. [11] employed MP2 ab-initio computations to evaluate the combustion enthalpies and heats of formation for 31 compounds. Although they have not reported the accuracy of their theoretically calculated combustion enthalpies in comparison with the experimental data, their theoretically reported values demonstrate significant deviations compared to the experiment, as we show in table 3.3 below. Audran and co-workers [12] employed ab-initio computations to calculate combustion enthalpies using four different levels of theory. They reported significant deviations between the predicted and experimentally determined data and proposed linear relationships to empirically improve the theoretically predicted combustion enthalpies. Mazzuca et al. [13] studied seven ab-initio methods for evaluating the combustion enthalpies of 31 compounds. To reduce the deviations they observed between the theoretically predicted and experimentally determined combustion enthalpies, they suggested empirically scaling the theoretically predicted combustion enthalpies with scaling factors ranging from 0.9846 to 1.1866, depending on the employed level of theory. Even with this empirical scaling, they could not achieve any mean unsigned errors better than 3%.

Considering that evaluation of enthalpies of chemical reactions by quantum mechanical computations is one of the most widely benchmarked and well-established applications of quantum chemistry, the present study investigates the reasons behind the reported deviations between the theoretically computed and experimentally determined enthalpies of combustion reactions. We provide insights into the appropriate treatment of the error sources and introduce guidelines for high accuracy determination of combustion enthalpies by theoretical computations.

3.2 Theoretical evaluation of combustion enthalpies

Precise evaluation of enthalpies of gas-phase reactions, as a trivial task in computational chemistry, is commonly achieved via ab-initio computation of enthalpies for individual molecules contributing to the reaction. Nevertheless, when it comes to combustion reactions, considering that the reactants and products might not always be in the gaseous state, phase

change thermodynamics can also play an important role and should be taken into account, which significantly adds to the computational challenges. In our literature survey, however, we noticed that the phase change enthalpies are commonly overlooked, which is one of the main contributors to the reported inconsistency between the theoretically evaluated and experimentally determined combustion enthalpies, as demonstrated in the results section.

For appropriate treatment of phase change thermodynamics in combustion reactions, careful attention should be paid to differences in defining the state of reactants and products and to different conventions in reporting combustion enthalpies. The experimentally determined combustion enthalpy data, commonly measured by oxygen bomb calorimetry, are typically reported either as gross or net heat of combustion. Gross heat of combustion refers to the total amount of heat released in the calorimetry experiment, where both the reactants and products are nearly at room temperature and in their standard states [14]. On contrast, for the net heat of combustion, while the reactants are considered in their standard states, the combustion products are assumed to be in the gas phase [14]. Clearly, the most obvious deviation between the reported gross and net heats of combustion is due to the heat released by the condensation of water molecules produced in the combustion reaction. It occurs as a result of cooling down the combustion products by the heat bath in which the combustion chamber is placed during the calorimetry experiment.

In addition to phase change thermodynamics, inaccuracies in theoretical computations, as well as experimentation can both significantly influence the accuracy of the obtained results. One of the most influencing parameters is the accuracy of the employed level of theory. Additionally, inappropriate energy minimization as the mandatory step before carrying out ab-initio calculation of thermal energy can remarkably reduce the accuracy of obtained results. The main aim of the present study is to investigate and benchmark the impact of such intricacies and demonstrate appropriate employment of ab-initio computations for high accuracy prediction of combustion enthalpy.

3.3 Computational details

The experimentally determined gross heats of combustion for 40 hydrocarbons reported by Walters [15] were used as reference combustion enthalpies. The full list of the studied hydrocarbons is reported in table 3.1. The selected hydrocarbons only contain C, H, and O atoms, to avoid complications e.g. due to solvation of nitric or sulfuric acid in water, produced by combustion of molecules containing nitrogen or sulfur and leading to contributions to the measured combustion enthalpy [14].

Computation of in vacuo enthalpies of compounds was carried out by normal mode analysis based on the rigid rotor harmonic oscillator approximation, as a standard approach for estimation of thermodynamic quantities in theoretical chemistry [16]. To that end, for each compound, we first optimized the three-dimensional geometries of the molecules in vacuo. These optimized structures were then used to calculate the ground-state electronic energies and normal mode vibrational frequencies required for calculating thermal effects.

Considering that the molecular configurations found at this stage by geometry optimization can yield a wide range of energies, appropriate geometry optimization plays a significant role in the ab-inito-evaluated molecular enthalpies and hence the resulting combustion enthalpy, as discussed in the results section and demonstrated in figure 3.2. Accordingly, trying to search for geometries yielding the global minimum on the potential energy surface should be attempted during geometry optimization. To that end, we considered multi-start geometry optimization using 20 different initial structures generated via the genetic algorithm module of the open babel toolbox [17]. The configuration which yielded the lowest energy after optimization was then used for the calculation of the combustion enthalpy.

Quantum-mechanical (QM) computations were carried out at the DSD-PBEP86-D3/ Def2QZVP level of theory, which is one of the most accurate methods for thermochemistry evaluation [18]. To study the influence of the employed level of theory on the accuracy of the obtained results, we also computed the molecular enthalpies at the B3LYP/6-311+G(2d,p) level of theory.

Considering that the theoretically computed enthalpies are obtained for molecules in vacuo, for non-gaseous reactants, the QM evaluated enthalpy of the reactants in their standard state was estimated by subtracting their heat of phase change from the standard state to the gas phase from the initially computed in vacuo enthalpies. To that end, the phase change enthalpies were taken from the NIST database. Similarly, considering that the reference data used in the present study are gross heats of combustion, the vaporization enthalpy of water with the value of 43.898 kJ/mol [19] was also subtracted from the QM evaluated enthalpy of water in vacuo to yield the QM enthalpy of water in the liquid state.

In the calculation of the enthalpy of O_2 molecules, we considered the triplet state as the ground electronic state, as conventionally employed in theoretical computation of combustion enthalpy [11]. At the DSD-PBEP86-D3/ Def2QZVP level of theory, our QM estimation of the ground state energies with zero-point energy included resulted in -150.129205 and -150.179060 Hartree per molecule energies for the singlet and triplet multiplicity states of O_2, respectively, which indeed implies the triplet state as the ground state of this molecule.

Since measurement of combustion enthalpies is commonly carried out under 30 bar pressure [14], we also investigated the pressure impacts on the enthalpy of the studied compounds. To that end, we exploited the following thermodynamic relationship:

$$\left(\frac{\partial H}{\partial P}\right)_T = V + T\left(\frac{\partial S}{\partial P}\right)_T = V - T\left(\frac{\partial V}{\partial T}\right)_P . \tag{3.1}$$

Considering that the changes in thermal expansion of solids and liquids for increasing the pressure from 1 to 30 bar is negligible, we only calculated the pressure impacts on the enthalpies of gaseous compounds. To that end, the molar volume of gaseous compounds and their derivative with respect to temperature as required by Eq. (3.1) were calculated via the Redlich–Kwong equation of state defined as [20]:

$$\begin{aligned} P &= \frac{RT}{V-b} - \frac{a}{\sqrt{T}\, V\,(V+b)}, \\ a &= 0.42748\,\frac{R^2 {T_c}^{2.5}}{P_c}, \\ b &= 0.08664\,\frac{R\,T_c}{P_c}, \end{aligned} \tag{3.2}$$

where R is the universal gas constant, V is the molar volume, and T_c and P_c are the critical temperature and pressure, respectively. Accordingly, for pressures from 1 to 30 bar with 1 bar intervals, the molar volumes were calculated for temperatures from 280K to 320K with 1K intervals via solving Eq. (3.2) by the bisection method. Using the calculated $V - T$ values for each pressure, a third-order polynomial was fitted and used to calculate the partial derivative $\frac{\partial V}{\partial T}$ as required by Eq. (3.1). Using the calculated molar volumes and $\frac{\partial V}{\partial T}$ for all pressures, the pressure-induced enthalpy changes were calculated by numerically approximating the following integral:

$$\Delta H = \int_1^{30} \left(V - T\left(\frac{\partial V}{\partial T}\right)_P\right) dp\,. \tag{3.3}$$

The accuracies of the predicted combustion enthalpies are reported as Average Absolute Deviation (AAD) and percentage Average Absolute Relative Error (MUE), defined as:

$$\mathrm{AAD} = \frac{1}{N}\sum\left(\left|y_i^{exp} - y_i^{pred}\right|\right), \tag{3.4}$$

$$\mathrm{AARE\%} = \frac{1}{N}\sum\left(\left|\frac{y_i^{exp} - y_i^{pred}}{y_i^{exp}}\right|\right)\times 100 \tag{3.5}$$

All the computations were carried out in Gaussian 16 software [21] on the High-Performance Computing center clusters of the Christian-Albrechts-University of Kiel.

3.4 Results and discussions

The details of the computed molecular enthalpies at the DSD-PBEP86-D3/ Def2QZVP level of theory based on the previously discussed recipes are reported in table 3.1, together with a direct comparison to the experiment. Also, for the same level of theory the calculated enthalpies for H_2O, CO_2, and O_2, as molecules involved in all combustion reactions are reported in table 3.2.

Using the QM-evaluated enthalpies corrected for phase change enthalpies of water and reactants, the predicted combustion enthalpies yielded AAD, AARE% and correlation coefficient of 17.68 kJ/mole, 0.623% and 0.99999, respectively. These results, which are directly calculated by ab-initio computation without any further correction, show a remarkable improvement compared to results reported elsewhere. For example, the theoretically calculated combustion enthalpies reported by Mazzuca et al. [13] yielded a AARE of roughly 3%, even after being empirically scaled. According to these results, the pressure and non-ideality impacts can alter the predictability of combustion enthalpy only slightly.

To further improve the accuracy of the theoretically computed combustion enthalpies, using the experimentally determined heats of combustion and QM predicted enthalpies of the reactants, we computed the optimum values of enthalpies for H_2O, CO_2, and O_2 molecules which yielded the minimum AAD. Comparing these values with the theoretically computed enthalpies reported in table 3.2 shows excellent agreements, with negligible percentage deviations.

Using the optimized enthalpies of H_2O, CO_2, and O_2 and QM evaluated values for the reactants, we could theoretically reproduce the experimentally determined combustion enthalpies with AAD, AARE% and correlation coefficient of 1.556 kJ/mole, 0.072% and 0.99999, respectively, which are the most accurate results ever reported for evaluation of combustion enthalpy, to the best of our knowledge. A graphical comparison of the

Table 3.1 Details of the theoretically calculated and experimental data.

Compound	Std.	$\Delta H_{std\text{-}gas}$	$H_{QM,\ reactant}$	$\Delta H_{comb.,QM}$	$\Delta H_{comb.,QM,opt,P}$	$\Delta H_{comb.,QM,opt}$	ΔH_{exp}
Oxyrane	g	0	-403132.59	-1312.7	-1305.6	-1305.54	-1305.53
Cyclopentane	l	28.8	-514947.83	-3307.8	-3288.88	-3288.78	-3288.85
Ethylbenzene	l	41	-814646.57	-4599.2	-4561.19	-4561.57	-4561.44
2-Butanone	l	34	-609289.96	-2457.59	-2442.73	-2442.82	-2442.95
Methanol	l	37.6	-303387.53	-727.72	-726.78	-726.76	-726.47
Cyclobutane	g	0	-411847.57	-2756.94	-2742.04	-2741.72	-2742.09
Acetone	l	31.27	-506296.27	-1800.15	-1789.07	-1789.19	-1789.6
Dimethyl ether	g	0	-406302.62	-1463.75	-1459.25	-1458.99	-1459.71
2-Propanol	l	45	-509403.24	-2014.26	-2005.76	-2005.69	-2004.92
Ethane	g	0	-209058.72	-1564.61	-1557.61	-1559.4	-1558.59
Acetaldehyde	g	0	-403245.3	-1199.99	-1192.91	-1192.83	-1191.93
Cyclopropane	g	0	-308851.28	-2102.1	-2090.96	-2090.69	-2089.79
Formic acid	l	46.3	-497680.78	-256.43	-253.19	-253.52	-254.46
Ethanol	l	42.3	-406394.38	-1371.99	-1367.28	-1367.23	-1366.23
Butane	g	0	-415038.88	-2886.7	-2874.42	-2873.89	-2874.96
Ethyl acetate	l	35	-806639.85	-2250.74	-2236.14	-2236.42	-2237.68
Isopropyl benzene	l	44	-917644.63	-5252.27	-5210.48	-5210.83	-5212.17
Diethyl ether	l	27.1	-612335.26	-2733.37	-2721.09	-2721	-2722.42
Benzene	l	33.9	-608647.31	-3296.2	-3265.76	-3266.18	-3264.75
1,4-Dioxane	l	38	-806515.02	-2375.57	-2360.98	-2361.25	-2362.73
1,2-Ethanediol	l	65	-603714.14	-1195.28	-1190.83	-1190.97	-1189.44
Phenol	s	69.7	-806003.58	-3082.98	-3052.81	-3053.41	-3051.84
vinyl acetate	l	37.2	-803470.61	-2098.9	-2081.72	-2082.18	-2080.62
Propanol	l	47	-509388.49	-2029.01	-2020.52	-2020.45	-2018.73
Heptane	l	36	-724039.74	-4839.22	-4815.32	-4814.99	-4813.15
Cyclohexane	l	33.1	-617968.58	-3938.18	-3915.48	-3915.35	-3917.19
1-pentanol	l	57	-715372.87	-3346.88	-3330.82	-3330.71	-3328.86
Glycerol	l	91.7	-904045.28	-1658.31	-1650.35	-1650.64	-1652.52
Propane	g	0	-312048.73	-2225.73	-2217.21	-2216.71	-2218.62
Acetic acid	l	50.3	-600705.65	-882.69	-875.67	-875.98	-874.05
Pentane	l	26.5	-518055.34	-3521.37	-3505.03	-3504.74	-3506.75
Isopropyl ether	l	32.26	-818344.07	-4026.81	-4006.96	-4006.83	-4008.9
Furan	l	27.71	-603001.12	-2104.27	-2084.24	-2084.7	-2082.44
Toluene	l	37	-711652.04	-3942.6	-3908.38	-3908.77	-3906.28
Hexane	l	31	-621049.89	-4177.95	-4157.83	-4157.52	-4160.07
1-Methylnaphthalene	l	59	-1114374	-5861.91	-5804.81	-5805.67	-5808.23
Benzaldehyde	l	48	-905857.39	-3559.21	-3522.68	-3523.44	-3526.08
Cyclohexene	l	33.57	-614808.73	-3776.95	-3751.67	-3751.73	-3748.59
1-Butene	g	0	-411869.34	-2735.16	-2720.3	-2719.94	-2715.74
m-Cresol	l	60	-908995.93	-3741.76	-3707.8	-3708.38	-3702.26

All enthalpies are in kJ/mole
The columns from left to right represent:
Std.: standard state (gas=g, liquid=l, solid=s)
$\Delta H_{std\text{-}gas}$: The enthalpy of phase change from the standard state to the gas phase
$H_{QM,reactant}$: The QM enthalpies of individual reactants in the gas phase
$\Delta H_{comb.,QM}$: Combustion enthalpy directly obtained via QM enthalpy of reaction
$\Delta H_{comb.,QM,opt,P}$: Combustion enthalpy corrected for pressure impacts using optimized enthalpies for H2O, CO2, and O2 and QM enthalpy of reactants
$\Delta H_{comb.,QM,opt}$: Combustion enthalpy without pressure correction using optimized enthalpies for H2O, CO2, and O2
ΔH_{exp}: The experimentally determined data

Table 3.2 QM calculated and optimum enthalpies of H_2O, CO_2, and O_2.

	H_2O (l)	CO_2(g)	O_2(g)
QM	-200464.123	-494616.132	-394286.085
Optimized	-200466.979	-494610.356	-394286.663

theoretically evaluated and experimentally determined combustion enthalpies is depicted in figure 3.1.

As can be inferred from the above-mentioned results, even the very slight deviations in the QM calculated enthalpies of H_2O, CO_2, and O_2 from the optimum values lead to an increase of 16.124 kJ/mol in the obtained AAD. This implies that the accuracy of the employed computational method plays a key role in the high-precision evaluation of the combustion enthalpies and should be carefully considered.

This can also be inferred from table 3.3, which provides a comparison of theoretically predicted combustion enthalpies for some combustion reactions in the gas phase reported in the literature with our results. These results clearly show large deviations between the results obtained via different levels of theory. While the computations at the MP2 level of theory typically yield more satisfactory results, other levels of theory clearly yield inappropriate estimations of molecular enthalpies.

To further demonstrate the importance of the applied level of theory, we also computed the combustion enthalpies at the B3LYP/6-311+G(2d,p) level of theory for the same dataset and computational details.

According to these results, for computations at B3LYP /6-311+G(2d,p) level of theory and after re-optimizing the enthalpies of H_2O, CO_2, and O_2, we obtained AAD and AARE% of 16.239 kJ/mol and 0.825%, respectively. Without optimizing the enthalpies of H_2O, CO_2, and O_2, the original QM computations yielded AAD and AARE% of 130.030 kJ/mol and 5.24%. Therefore, the results obtained via the B3LYP /6-311+G(2d,p) level of theory are roughly one order of magnitude less accurate than those obtained via DSD-PBEP86-D3/ Def2QZVP level of theory.

By analyzing the details of computed energies we noticed that molecular thermal energies, i.e. the kinetic energy due to rotation and translation energy and vibrational energies, contribute on average only 0.625 % and 0.541% to the computed combustion enthalpies, while the changes in ground state electronic energies of reactants and products are the main contributions to the heat released by combustion. Accordingly, the accuracy of the employed level of theory in reproducing the ground state electronic energy and not the thermal effects play a key role in the accuracy of the obtained results. By comparing the thermal and

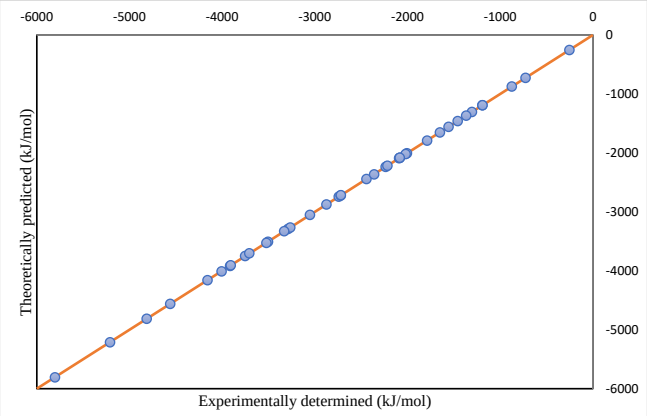


Fig. 3.1 comparison of theoretically predicted and experimentally determined combustion enthalpies.

electronic energies evaluated via DSD-PBEP86-D3/ Def2QZVP and B3LYP /6-311+G(2d,p) levels of theory, we observed an AAD of 148.425 kJ/mol between the computed ground state electronic energies, while for thermal energies the AAD was only 0.781 kJ/mol. This is because the theoretical methods are usually parameterized using the quantities which are related to the thermal energies, such as vibrational frequencies or thermodynamic quantities, while the absolute values of ground-state energies are not experimentally measurable. These results also reveal why the accuracy of theoretical methods for combustion reactions is so different from the benchmark results obtained for other case studies. The reason is that the large amount of energy released by combustion reactions is mainly due to electronic energies, which implies substantial deviations between electronic energies of reactants and products.

The DSD-PBEP86-D3/Def2QZVP level of theory used in the present study supersedes most of the conventionally accepted functionals in studying thermochemistry and for reproducing the thermal effects [22]. However, in terms of accuracy for reproducing ground state electronic energy, this method still has slight inaccuracies compared to the high level and computationally demanding methods such as CCSD(T), which is commonly considered as the gold standard in theoretical chemistry [23].

Based on all these observations, for the most reliable evaluation of combustion enthalpies in which a substantial difference between the ground state energies of reactants and products exists, employing methods which are specifically parameterized for reproducing ground state electronic energies like CCSD or DLPNO-CCSD(T), which is a cost-effective method for accurate reproduction of CCSD(T) energies [24] for estimating ground state energies, and using other common methods for evaluating the thermal effects is recommended.

After the accuracy of the employed level of theory, the second most important source of inaccuracy in theoretically evaluated combustion enthalpies arises from inappropriately

Table 3.3 Comparison of QM evaluated enthalpies for different levels of theory.

Compound	method	source	$\Delta H_{comb,QM,opt,gas}$ (kJ/mole) *	ΔH_{exp}(kJ/mole)**
Ethane	PBEP86-D3/Def2QZVP	Present study	-1427.915	-1426.895
	MP2	[11]	-1422.7	
	B3LYP/6-311G(d,p)	[12]	-1284.36	
	TPSS-TPSS/6-311G+ +(df,pd)	[12]	-1275.44	
Propane	PBEP86-D3/Def2QZVP	Present study	-2041.343	-2043.304
	MP2	[11]	-2036.6	
	B3LYP/6-311G(d,p)	[12]	-1853.63	
	TPSS-TPSS/6-311G+ +(df,pd)	[12]	-1839.42	
n-Butane	PBEP86-D3/Def2QZVP	Present study	-2654.639	-2655.470
	MP2	[11]	-2650.2	
	B3LYP/6-311G(d,p)	[12]	-2423.40	
	TPSS-TPSS/6-311G+ +(df,pd)	[12]	-2403.74	
Cyclopropane	PBEP86-D3/Def2QZVP	Present study	-1959.477	-1958.092
	MP2	[11]	-1957.2	
Cyclobutane	PBEP86-D3/Def2QZVP	Present study	-2566.201	-2566.498
	MP2	[11]	-2566.0	
Cyclohexane	PBEP86-D3/Def2QZVP	Present study	-3685.178	-3686.904
	MP2	[11]	-3685.0	
n-hexane	PBEP86-D3/Def2QZVP	Present study	-3881.517	-3883.785
	HF/6-311+ +G(3df,3pd)	[12]	-3389.77	
	B3LYP/6-311G(d,p)	[12]	-3563.42	
	B3LYP/6-311+ +G(3df,3pd)	[12]	-3765.90	
	B3LYP/6-311+ +G(3df,3pd)	[12]	-3737.44	
	PBEPBE/6-311+ +G(3df,3pd)	[12]	-3654.47	
	TPSS-TPSS/6-311G+ +(df,pd)	[12]	-3532.84	
1-pentanol	PBEP86-D3/Def2QZVP	Present study	-3124.418	-3122.477
	HF	[13]	-2688.45	
	MP2	[13]	-3189.30	
	B3LYP	[13]	-2963.65	
	M06	[13]	-3286.41	
	ωB97X-D	[13]	-2939.76	

* Considering that the reported values in the literature are all computed for the gas phase reactions, the values we report here also are computed considering all the reactants and products in the gas phase and are not corrected for phase change enthalpies.

** Since the theoretically determined data here are computed for the gas phase, ΔH_{exp} data are also corrected to be the heat of combustion for the gas phase via subtracting the contributions of phase change enthalpies of reactant and products from the gross heats of combustion reported in table 1. It should however be noted that in references 12 and 13, the theoretically predicted values are directly compared with the gross heats of combustion which resulted in greater inconsistencies.

optimized structures with high energies. As for almost all polyatomic molecules, several local minima exist on the potential energy surface, geometry optimizations started from different initial structures can result in quite diverse geometries and molecular energies and consequently different predicted combustion enthalpies. As an example, theoretical computations on the two structures of acetic acid depicted in figure 3.2, which were obtained by geometry optimizations started from different initial structures, yield quite different combustion enthalpies. While QM computations for the low-energy structure (without optimizing the enthalpies of H_2O, CO_2, and O_2 yields a combustion enthalpy with 8.64 kJ/mole absolute error, the same computation for the high-energy structure results in 29.76 kJ/mole absolute error.

Inaccuracies from inappropriately optimized structures can be avoided by employing efficient general global optimization algorithms [25-27] or rotamer searches [28] or, for small molecules as those considered here, using multi-start optimization. In the present study, we selected the latter approach as discussed in the previous section.

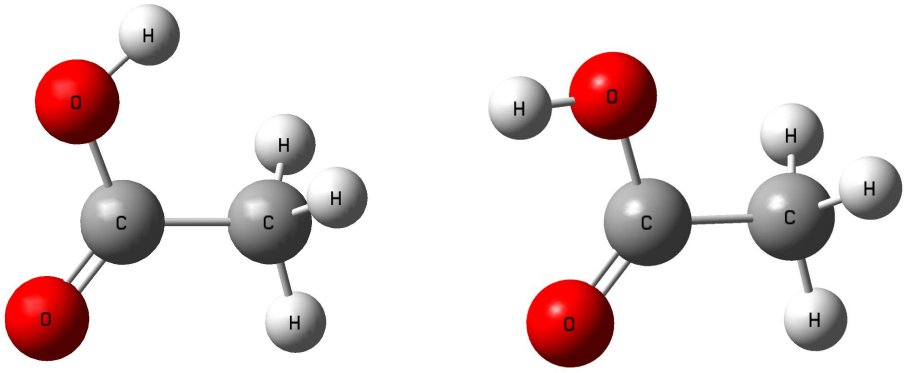


Fig. 3.2 Two locally optimized structures of the acetic acid yield gas phase QM calculated enthalpies of -600634.230 kJ/mole (left) and -600655.349 kJ/mole (right).

Yet another reason for deviation between the QM predicted and optimum enthalpies can be overlooking non-ideality effects. As discussed earlier, increasing the ambient pressure can directly influence the phase change and gas-phase enthalpies, while QM enthalpies are computed for molecules in vacuo. We studied this direct impact of pressure on gas-phase enthalpies via Eq. (3.3). However, this correction could only slightly improve the accuracy of predicted combustion enthalpies, as can be seen in table 3.1. The more significant impact of the ambient pressure on gas-phase enthalpies can be attributed to the formation of molecular clusters in the gas phase at high pressures. For example, for the accurate evaluation of the phase change enthalpy as well as the saturation vapor pressure of water, it has been shown that clustering of two or further molecules in the gas phase should be taken into account

[29]. Such gas phase clustering reduces the gas phase enthalpy compared to the in vacuo state, which is in line with the observed difference between the QM evaluated and optimized enthalpy of water, reported in table 3.2.

In addition to the inaccuracies resulting from theoretical computations, systematic or operational errors in experimental data can also contribute to inconsistency between the theoretically evaluated and reference data. For example, we observed 1.59 kJ/mol average absolute deviation in phase change enthalpies of our studied reactants between the NIST and DIPPR databases which results in the same deviation between the theoretically predicted gross combustion enthalpies calculated using each one of these two databases. Similar to the vaporization enthalpy, the experimentally determined combustion enthalpies also show some variations from different sources. For example, slight inaccuracy in measuring the combustion enthalpy of benzoic acid, which is used to calibrate the calorimeter [14], can result in a linearly distributed deviation among measured combustion enthalpies of all other compounds. That can be a potential reason for the suitability of a linear curve fitting to empirically correct the predicted combustion enthalpies, proposed in several studies as discussed earlier.

3.5 Summary and conclusion

In summary, in the present study, we have discussed computational details which can result in a high-accuracy evaluation of combustion enthalpy. To that end, the main considerations in theoretical computations should be directed towards selecting an appropriate level of theory and searching for structures with energies close to the global minimum energy, e.g. by multi-start optimization. In reproducing the net heat of combustion, the phase change enthalpy of the reactants should be subtracted from the QM-evaluated gas-phase enthalpies. For the gross heat of combustion, the vaporization enthalpy of water should also be subtracted from the QM-evaluated gas-phase enthalpy of water. Accordingly, the inaccuracies in the mentioned phase change enthalpies, as well as the experimental measurement of combustion enthalpy, can also contribute to inconsistencies between the theoretically predicted and experimentally determined combustion enthalpies, as demonstrated in the present study.

Acknowledgment

The author wish to thank Bernd Hartke in CAU university of Kiel for his fruitful discussions and reviewing this work.

References:

1. King, M. K.; Ross, H. D., Overview of the NASA microgravity combustion program. AIAA journal 1998, 36 (8), 1337-1345.
2. Suleyman A. Gokoglu, D. L. D., Dennis P. Stocker, Paul V. Ferkul, Sandy L. Olson, Michael C. Hicks A Researcher's Guide to: Combustion Science. NASA 2016.
3. Motil, B.; Urban, D., Combustion, Complex Fluids, and Fluid Physics Experiments on the ISS. 2012.
4. Zhu, Y.; Zhou, C.-W., Chemical kinetics study of 1, 3-butadiene+ HO2; implications for combustion modeling and simulation. Combustion and Flame 2020, 221, 241-255.
5. Kopp, W. A.; Kröger, L. C.; Döntgen, M.; Jacobs, S.; Burke, U.; Curran, H. J.; Heufer, K. A.; Leonhard, K., Detailed kinetic modeling of dimethoxymethane. Part I: Ab initio thermochemistry and kinetics predictions for key reactions. Combustion and Flame 2018, 189, 433-442.
6. Ye, L.; Zhang, L.; Qi, F., Ab initio kinetics on low temperature oxidation of iso-pentane: the first oxygen addition. Combustion and Flame 2018, 190, 119-132.
7. Shi, X.; Wang, Q.; Violi, A., Chemical pathways for the formation of benzofuran and dibenzofuran in combustion. Combustion and Flame 2020, 212, 216-233.
8. Shyamala, B.; Lal, S.; Chowdhury, A.; Namboothiri, I. N.; Kumbhakarna, N., Cubane decomposition pathways–A comprehensive study. Combustion and Flame 2018, 197, 111-119.
9. Osmont, A.; Catoire, L.; Gökalp, I.; Yang, V., Ab initio quantum chemical predictions of enthalpies of formation, heat capacities, and entropies of gas-phase energetic compounds. Combustion and Flame 2007, 151 (1-2), 262-273.
10. Swihart, M. T.; Catoire, L., Thermochemistry of aluminum species for combustion modeling from ab initio molecular orbital calculations. Combustion and flame 2000, 121 (1-2), 210-222.
11. Whyman, G.; Savoskin, M.; Yaroshenko, A.; Kapkan, L.; Popov, A., Straightforward ab initio calculation of enthalpies of combustion and formation of hydrocarbons. Journal of Molecular Structure: THEOCHEM 2003, 637 (1-3), 183-187.
12. Audran, G.; Marque, S. R.; Siri, D.; Santelli, M., Enthalpy of Combustion on n-Alkanes. Quantum Chemical Calculations up to n-C60H122 and Power Law Distributions. ChemistrySelect 2018, 3 (31), 9113-9120.
13. Mazzuca, J. W.; Downing, A. R.; Potter, C., Empirically corrected electronic structure calculations applied to the enthalpy of combustion physical chemistry laboratory. Journal of Chemical Education 2019, 96 (6), 1165-1170.
14. D4809-13, A., Standard Test Method for Heat of Combustion of Liquid Hydrocarbon Fuels by Bomb Calorimeter (Precision Method). ASTM International 2013.
15. Walters, R. N., Molar group contributions to the heat of combustion. Fire and materials 2002, 26 (3), 131-145.

16. McQuarrie, D. A., Statistical thermodynamics. 1973.
17. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. Journal of cheminformatics 2011, 3 (1), 1-14.
18. Kozuch, S.; Martin, J. M., DSD-PBEP86: in search of the best double-hybrid DFT with spin-component scaled MP2 and dispersion corrections. Physical Chemistry Chemical Physics 2011, 13 (45), 20104-20107.
19. Hui, A. X. H., ENTHALPIES OF VAPORIZATION OF SOME MUITICHLORO-ALKANES. Acta Physico-chimica Sinica 1989, 05.
20. Murdock, J. W., Fundamental fluid mechanics for the practicing engineer. CRC Press: 2018.
21. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., Gaussian 16. Revision A 2016, 3.
22. Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S., A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. Physical Chemistry Chemical Physics 2017, 19 (48), 32184-32215.
23. Kesharwani, M. K.; Brauer, B.; Martin, J. M., Frequency and zero-point vibrational energy scale factors for double-hybrid density functionals (and other selected methods): can anharmonic force fields be avoided? The Journal of Physical Chemistry A 2014, 119 (9), 1701-1714.
24. Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F., Natural triple excitations in local coupled cluster calculations with pair natural orbitals. The Journal of chemical physics 2013, 139 (13), 134101.
25. Dieterich, J. M.; Hartke, B., Error-safe, portable, and efficient evolutionary algorithms implementation with high scalability. Journal of chemical theory and computation 2016, 12 (10), 5226-5233.
26. Hartke, B., Wiley Interdiscip. Rev.: Comput. Mol. Sci 2011, 1, 879-887.
27. Dieterich, J. M.; Hartke, B., OGOLEM: Global cluster structure optimisation for arbitrary mixtures of flexible molecules. A multiscaling, object-oriented approach. Molecular Physics 2010, 108 (3-4), 279-291.
28. Pracht, P.; Bohle, F.; Grimme, S., Automated exploration of the low-energy chemical space with fast quantum chemical methods. Physical Chemistry Chemical Physics 2020, 22 (14), 7169-7192.
29. Parsafar, G. A.; Moeini, V.; NAJAFI, B., Pressure dependence of liquid vapor pressure: an improved gibbs prediction. 2001.

Chapter 4

Theoretical evaluation of equilibrium constant for isotope exchange reactions

Project overview and motivation:

The equilibrium constant of boron isotope fraction between boric acid and borate in pure and saline water is of high geochemical importance. This parameter is extensively required to reconstruct the ancient seawater pH and atmospheric pCO2 via analysis of isotopic composition in fossils of marine shells. Theoretical calculation of this parameter has been aimed in several studies which resulted in reporting a wide range of inconsistent estimations and therefore, is still a controversial topic in geochemistry. It motivated us to consider this case study both for method development applicable to the theoretical evaluation of thermochemistry in solution and to provide a reliable estimation of the required equilibrium constant of boron isotope fractionation. We developed novel theoretical methods which can be used for achieving high accuracy results for a reduced computational cost, not only for isotope exchange reactions but also for other applications as well.

Novelty aspects:

- The proposed *partial* normal mode analysis and Boltzmann weighted averaging which allow cost-effective studying chemical reactions in the liquid phase by ab initio normal mode analysis and explicit solvent
- The employed level of theory which is the highest studied level compared to other works carried out for the same purpose, to date

- The number of studied configurations and cluster sizes are the largest, compared to other previously published reports
- Employing implicit solvent approaches and path integral molecular dynamics simulation to study isotope exchange reactions in *saline* water has not been studied elsewhere, to date.

Connection to other chapters:

Evaluation of the equilibrium constant by normal mode analysis is based on the knowledge acquired through studying combustion enthalpies with the same method in chapter 3. The methods which are developed based on employing the implicit solvent approach are cornerstones of the more advanced implicit solvent models proposed in chapters 8 and 9 as well as the thermodynamically effective molecular surfaces introduced in chapter 5. Furthermore, the cavity surfaces based on implicit solvation models are major inputs of machine learning models proposed in chapter 7 for estimation of vaporization enthalpy and its temperature dependence.

Contributions:

Conception and developing the *partial* normal mode analysis, revised Bigeleisen and Mayer approach, and Boltzmann weighted averaging methods and major contributor to the other methods developed in the context of this project, carrying out all the computations, major contribution in writing the manuscript.

Publication status:

This work is in the submission process and its preprint is available online (DOI:10.33774/chemrxiv-2021-lpdnx).

Theoretical evaluation of equilibrium constant for boron isotopes exchange between boric acid and borate in pure and saline water

Amin Alibakhshi[1,*], Julien Steffen[1], Carlos Pinilla[2,3], Bernd Hartke[1]

[1]Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, 24118 Kiel, Germany

[2]Departamento de Física y Geociencias, Universidad del Norte, Km 5 via Puerto Colombia, Colombia

[3]School of Chemistry, University of Bristol, Cantock's Close Road, BS8 1TS, Bristol, United Kingdom

Corresponding author: alibakhshi@pctc.uni-kiel.de

Abstract

Accurate estimation of the equilibrium constant of boron isotope fractionation between boric acid and borate (k_{3-4}) in water is of high geochemical importance, due to its contribution in reconstruction of ancient seawater pH and atmospheric CO_2. As a result, precise evaluation of k_{3-4} has been the subject of numerous studies, yielding diverse and controversial results. In the present study, we provide our estimation of k_{3-4}, theoretically evaluated via three different rigorous and high-precision approaches. Based on our results, we propose the most probable estimation of k_{3-4} within 1.028 to 1.030 for both pure and saline water.

4.1 Introduction

Studying equilibrium between boron species dissolved in seawater is of high geochemical importance, due to its contribution to the alkalinity and buffering capacity of seawater and thus its ability to mitigate pH changes, which is of vital importance for marine ecosystems [1,2].

Additionally, the equilibrium of boron species in seawater also plays a key role in studying and reconstruction of ancient seawater pH and atmospheric pCO_2 via analysis of boron isotopic ratios in marine shell fossils [3-7]. To that end, an accurately determined equilibrium constant of the following isotope exchange reaction between aqueous boric acid and borate (k_{3-4}):

$$^{10}B(OH)_3 + {}^{11}B(OH)_4{}^- \leftrightarrows {}^{11}B(OH)_3 + {}^{10}B(OH)_4{}^-$$

is of central importance and has been the subject of numerous studies. As the other more recent applications of studying the equilibrium between boron isotopes in water, we can refer

to the identification of nitrate pollution sources [8] and elucidation of shale weathering in the critical zone [9].

The earliest estimation of k_{3-4} was theoretically calculated by Kakihana et al. [10] who reported a value of 1.0194 for this equilibrium constant. They obtained this value through the partition functions they had computed based on an empirical valence force field. This initially reported value, however, has been widely criticized in several more recent studies which suggest larger values for k_{3-4} ranging from 1.025 to 1.035 [11-15,16]. It can especially be important because even the slight deviation between the initially reported estimation of k_{3-4} and the more recent values can have significant geochemical importance, as it results in a value for the estimated pH of the ocean that is almost one unit larger [16].

Among the recent evaluations of k_{3-4}, only one work reports experimentally determined values of 1.0308 ± 0.0023 and 1.0272 ± 0.0003 obtained for pure and seawater, respectively [15]. Considering that experimental measurement of isotope fractionation in isotope exchange reactions is typically highly sensitive to experimentation, theoretical methods are more commonly employed for this purpose and are required not only to evaluate isotope fractionation but also to support experimentally determined results.

Despite the importance of precise determination of k_{3-4}, our careful review of the theoretical works previously carried out for this purpose reveals a number of significant limitations in those works which might affect the accuracy of those reported results.

First and foremost, the employed theoretical approach for evaluation of k_{3-4} in all of the previously reported results are mainly limited to Normal Mode vibrational frequency Analysis (NMA) based on the harmonic oscillator approximation and thus neglect anharmonicity effects. The anharmonicity can play a significant role in isotope exchange reactions especially in light elements like boron [17].

The other limiting factor is the expected inaccuracy inherent in the employed theoretical levels in the previously reported results. The highest levels of theory employed in the previous studies are mainly limited to HF/6-31G(d) and B3LYP/6-31G(d) [13], HF/6-31G* followed by empirical scaling of frequencies [14] and PBE96 [18], which are not among those accepted as high accuracy methods for thermochemistry studies via vibrational frequency analysis [19,20]. Rustad and co-workers reported a calculation of k_{3-4} at the MP2/aug-cc-pVTZ level of theory [16] which was the highest employed level of theory for this purpose so far. However, these computations were carried out for clusters with only 11 water molecules, yielding k_{3-4} equal to 1.033. To empirically correct the inefficiency due to the small size of their studied clusters, they reported extrapolated values according to the pattern they observed for lower-level computations, which resulted in an estimation of k_{3-4} between 1.026 to 1.028.

Note that both the applied level of theory and the cluster size can significantly impact the theoretically evaluated k_{3-4}, as shown by Rustad et al. [16]. It implies that the small size of the studied clusters is another shortcoming in the previously reported estimations of k_{3-4}. The largest cluster size studied so far includes 34 water molecules [14], which is obviously below an optimal size to properly include long-range interactions.

Another limitation of the studies carried out previously is the limited number of configurations considered for calculating k_{3-4}. As for larger molecules containing many atoms, typically several local minima exist, geometry optimizations required for the NMA can yield quite diverse geometries with the possibility of very different energies, properties, and solute-solvent interactions for each one. This necessitates employing multiple molecular configurations for the studied clusters. The work of Rustad and co-workers [16] used 10 different solute-solvent configurations, but the results reported in other studies are based on only one configuration.

Finally, the currently reported theoretical studies of k_{3-4} are only limited to calculated equilibrium constants in pure water, while the practical applications of equilibrium of boron isotopes mainly require estimations of k_{3-4} for saline water.

The main aim of the present study is to address the abovementioned limitations and provide a reliable and highly precise theoretical estimation of k_{3-4} for both pure and saline water. To that end, we employ vibrational frequency NMA for 60 configurations of large clusters of boric acid and borate dissolved in pure water employing a high level of theory. Nevertheless, NMA with the explicit solvent approach cannot be conveniently employed for studying multi-component solvents since achieving appropriate sampling of a canonical ensemble in this way is challenging. Consequently, theoretical evaluation of k_{3-4} in saline water is achieved via NMA with the implicit solvent approach [21] as well as path integral molecular dynamics. The latter method also allows taking into account the anharmonicity impacts on the calculated partition functions [17].

4.2 Theory

4.2.1 Evaluation of equilibrium constants via NMA

The equilibrium constants of chemical reactions can be theoretically calculated via:

$$K_{eq} = \exp\left(-\frac{\Delta A}{k_B T}\right), \tag{4.1}$$

Table 4.1 The partition functions obtained based on rigid rotor harmonic oscillator approximation.

$Q_{\text{trans.}} = \left(\frac{2\pi M k_B T}{h^2}\right)^{\frac{3}{2}} V$
$Q_{\text{rot.}} = \frac{\sqrt{\pi}}{s}\left(\frac{8\pi^2 I_A k_B T}{h^2}\right)^{\frac{1}{2}}\left(\frac{8\pi^2 I_B k_B T}{h^2}\right)^{\frac{1}{2}}\left(\frac{8\pi^2 I_C k_B T}{h^2}\right)^{\frac{1}{2}}$ (nonlinear molecule)
$Q_{\text{rot.}} = \frac{8\pi^2 I k_B T}{s h^2}$ (linear molecule)
$Q_{\text{vib.}} = \frac{\exp\left(-\frac{hc\nu_i}{2k_B T}\right)}{1-\exp\left(-\frac{-hc\nu_i}{k_B T}\right)}$

M is the mass of the molecule, V is the volume of the system, h is the Planck constant, s is the symmetry number, I is the principial moment of inertia, c is the light speed (cm s^{-1}) and ν_i is the i'th normal mode harmonic vibrational frequency (cm^{-1}).

in which ΔA is the free energy change of the reaction, k_B is the Boltzmann constant and T is the temperature. Using the statistical thermodynamics definition of free energy which relates it to the partition function (Q) via $A = -k_B T \ln(Q)$, we can rewrite Eq. 4.1 as:

$$K_{eq} = \frac{\prod_i Q_{\text{product},i}}{\prod_i Q_{\text{reactant},i}}. \tag{4.2}$$

The total energy for each one of the reactants and products is traditionally split into contributions from translation, rotation, vibration and electronic energy, resulting in [22]:

$$Q = \sum \exp\left(-\frac{\varepsilon_{\text{trans.}} + \varepsilon_{\text{rot.}} + \varepsilon_{\text{vib.}} + \varepsilon_{\text{elect.}}}{k_B T}\right) = Q_{\text{trans.}} Q_{\text{rot.}} Q_{\text{vib.}} Q_{\text{elect.}}. \tag{4.3}$$

For isotope exchange reactions, the electronic partition functions of isotopomers cancel out as they are electronically the same and differ only by their nuclear masses. As a result, one only needs to compute the translational, rotational, and vibrational partition functions which are conventionally obtained based on the Rigid Rotor Harmonic Oscillator (RRHO) approximation reported in table 4.1 [22], and for nonlinear molecules results in:

$$Q = \alpha M^{\frac{3}{2}} \frac{(I_A I_B I_C)^{\frac{1}{2}}}{s} \prod_i \frac{\exp\left(-\frac{hc\nu_i}{2k_B T}\right)}{1-\exp\left(-\frac{-hc\nu_i}{k_B T}\right)}. \tag{4.4}$$

Here α is a constant which depends only on the volume and temperature of the system and cancels out for reactions under constant volume and temperature.

For an isotope exchange reaction *A+B $\leftrightarrows$A+*B (the heavier isotopomer is specified with *), Eq. (4.4) can be further simplified using the Teller-Redlich product rule, which for two isotopomers 1 and 2 implies:

$$\prod_i \frac{\nu_{i,2}}{\nu_{i,1}} = \left(\frac{M_2}{M_1}\right)^{\frac{3}{2}} \left(\frac{I_{A,2}I_{B,2}I_{C,2}}{I_{A,1}I_{B,1}I_{C,1}}\right)^{\frac{1}{2}} \prod_j \left(\frac{m_{1,j}}{m_{2,j}}\right)^{\frac{3}{2}} \tag{4.5}$$

where $m_{i,j}$ is the mass of jth atom in the isotopomer i. In what follows, for an isotope exchange reaction, combining Eqs. (4.4) and (4.5) yields:

$$K_{A-B} = \frac{\mathrm{RPFR}_B}{\mathrm{RPFR}_A}, \tag{4.6}$$

where RPFR is known as the reduced partition function ratio of the solute and is calculated via:

$$\mathrm{RPFR} = \prod_i \frac{\nu_i,*}{\nu_i} \left(\frac{\exp\left(-\frac{hc\nu_{i,*}}{2k_BT}\right)}{\exp\left(-\frac{hc\nu_i}{2k_BT}\right)}\right)\left(\frac{1-\exp\left(-\frac{hc\nu_i}{k_BT}\right)}{1-\exp\left(-\frac{hc\nu_{i,*}}{k_BT}\right)}\right). \tag{4.7}$$

Calculation of equilibrium constants of isotope exchange reactions via RPFR was first proposed by Bigeleisen and Mayer [23] and has been widely used to study isotope exchange reactions [24-31]. The main advantage of the Bigeleisen and Mayer approach is that it allows calculating the isotope exchange equilibrium constants solely via normal mode vibrational frequencies. Considering that those vibrational frequencies can be obtained also experimentally and therefore without requiring calculations of principal moments of inertia for the molecules, the Bigeleisen and Mayer approach provides a practical way to estimate the required equilibrium constants.

In the present study, in addition to the original Bigeleisen and Mayer approach, we also introduce and investigate a revision in the original implementation of the Bigeleisen and Mayer method in section 4.2.3 which allows achieving a more rigorous integration of the results obtained for the multiple molecular configurations.

4.2.2 Cost effective NMA with explicit solvent via partial normal mode analysis

Normal mode vibrational frequencies required by NMA are conventionally computed based on the RRHO approximation. To that end, the first derivative of the system Hamiltonian with respect to molecular geometry, i.e. the net force on each atom, must be zero, which necessitates geometry optimization as the first step.

For NMA in solution, taking into account the solvent effects is commonly achieved via two ways, namely implicit and explicit solvent approaches [21]. While in the explicit solvent

approach, the solute is placed in the cluster of a number of solvent molecules, in the implicit solvent approach the solute is placed in the cavity of an implicitly defined solvent, instead.

For NMA with explicit solvent, the required geometry optimization is commonly applied for the whole solute-solvent cluster. This, however, makes the computations far more challenging and for high theoretical levels or large cluster sizes sometimes too expensive to be affordable. This is the reason for commonly limiting those computations to a very small number of solute-solvent configurations, small cluster sizes, or low computational levels, as discussed earlier. More importantly, geometry optimization of the whole cluster can result in optimized geometries with an unpredictable density, sometimes inconsistent with that of the real studied system.

To overcome all these limitations, we introduce here the partial normal mode analysis. According to this approach, for multiple molecular configurations of solute and solvent clusters with a constant density, we carry out geometry optimization and NMA only for the solute while the solvent molecules are kept frozen during the optimization and NMA. Employing partial normal mode analysis followed by the thermodynamically rigorous integration, introduced in section 4.2.3, allows us to not only reduce the computational costs and challenges by several orders of magnitude but also to achieve accurate results due to studying clusters with a correct density and being able to study contributions from a higher number of configurations and for larger cluster sizes. Partial normal mode analysis for multiple molecular configurations also allows capturing anharmonic effects in the obtained results due to carrying out the NMA for diverse computed frequencies and configurations.

4.2.3 Integrating multiple configuration results

As discussed earlier, appropriate evaluation of thermochemistry in solution with explicit solvents requires considering contributions from multiple solute-solvent configurations. The results obtained for each configuration then should be integrated to yield the required quantities. To that end, the most convenient integration approach is obviously averaging the results obtained for all configurations. It has been the method of choice to integrate the results of 10 configurations by Rustad and co-workers [16]. For extremely large numbers of appropriately sampled configurations, this averaging yields the exact estimation of the thermodynamic properties. Nevertheless, integrating smaller numbers of configurations requires going beyond the simple arithmetic mean and employing a more rigorous integration. For this purpose, here we propose a revision of the Bigeleisen and Mayer approach which allows exploiting Boltzmann-weighted averaging, as introduced in the following.

Considering the probability of the existence of any observable in the canonical ensemble, which is proportional to $e^{-\frac{H}{k_B T}}$ based on the Boltzmann statistics where H is the Hamiltonian of the molecule [32], we suggest a more rigorous Boltzmann weighted averaging defined as:

$$K_{eq} = \frac{\prod_i \left\langle Q_{\text{product},i} \right\rangle}{\prod_i \left\langle Q_{\text{reactant},i} \right\rangle}, \tag{4.8}$$

where $\langle Q \rangle$ is an ensemble-averaged partition function and is computed via:

$$\langle Q \rangle = \frac{\sum_i Q_i \exp\left(-\frac{\varepsilon_i}{k_B T}\right)}{\sum_i \exp\left(-\frac{\varepsilon_i}{k_B T}\right)}. \tag{4.9}$$

Here, Q_i and ε_i are the calculated partition function and energy of the solute for configuration i.

To implement the Boltzmann-weighted averaging in an approach similar to the Bigeleisen and Mayer method, we propose fractionizing the Teller-Redlich product rule defined in Eq. (4.5) for isotopomer a as follows:

$$\prod_i \nu_{i,a} = \alpha'(M_a)^{\frac{3}{2}} (I_{A,a} I_{B,a} I_{C,a})^{\frac{1}{2}} \prod_j \left(\frac{1}{m_{a,j}}\right)^{\frac{3}{2}}, \tag{4.10}$$

where α' is a coefficient which is constant for isotopomers of a molecule. If we substitute Eq. (4.10) in Eq. (4.4), we obtain:

$$Q = \frac{\alpha}{\alpha' s} \prod_j (m_{a,j})^{\frac{3}{2}} \prod_i \nu_i \frac{\exp\left(-\frac{hc\nu_i}{2k_B T}\right)}{1 - \exp\left(-\frac{hc\nu_i}{k_B T}\right)}. \tag{4.11}$$

Knowing that in calculating equilibrium constants for isotope exchange reactions, the factor $\frac{\alpha}{\alpha'} \prod_j (m_{a,j})^{\frac{3}{2}}$ will be cancelled out, we can now define the Reduced Partition Function (RPF) as:

$$\text{RPF} = \frac{\alpha' Q}{\alpha \prod_j (m_{a,j})^{\frac{3}{2}}} = \frac{1}{2} \prod_i \nu_i \frac{\exp\left(-\frac{hc\nu_i}{2k_B T}\right)}{1 - \exp\left(-\frac{hc\nu_i}{k_B T}\right)}. \tag{4.12}$$

Using RPF, the equilibrium constant is then calculated via:

$$K_{eq} = \frac{\prod_i \left\langle \text{RPF}_{\text{product},i} \right\rangle}{\prod_i \left\langle \text{RPF}_{\text{reactant},i} \right\rangle}. \tag{4.13}$$

Here, $\langle \text{RPF} \rangle$ is the ensemble average of the reduced partition function and is calculated via RPF_i and ε_i which are the RPF and energy of the solute in configuration i as follows:

$$\langle \mathrm{RPF} \rangle = \frac{\sum_i \mathrm{RPF}_i \exp\left(-\frac{\varepsilon_i}{k_B T}\right)}{\sum_i \exp\left(-\frac{\varepsilon_i}{k_B T}\right)}. \tag{4.14}$$

The energy of solute molecules required by Eq. (4.14) can be obtained most straightforwardly using the implicit solvent approaches. To that end, we employed the IEF-PCM continuum solvation model for the same level of theory to evaluate the in-solution energy of various conformers of the solute obtained by geometry optimization in each cluster.

Considering that the ε_i values are total ground state electronic energies and thus large negative numbers, to avoid numerical blow-up we employed subtracting the minimum of obtained ε_i values from all of them. This is equivalent to multiply both the denominator and numerator of Eq. (4.14) by $e^{-\frac{\min(\varepsilon_i)}{k_B T}}$ and therefore allows numerical calculation of the reduced partition functions without losing accuracy.

4.2.4 NMA with implicit solvent

Although the NMA with explicit solvent possesses a number of obvious advantages, e.g. taking into account intricate solute-solvent interactions such as hydrogen bonding, at the same time it suffers from some limitations. In addition to the challenges discussed in the previous section, the computational costs even for the partial normal mode analysis can still be quite demanding, especially for large clusters and high levels of theory. The other main limitation of NMA with explicit solvent is the appropriate treatment of multi-component solvents like saline water and the challenges of correctly positioning the ions.
To overcome such limitations, for evaluation of k_{3-4} in saline water, we exploited NMA with implicitly defined solvent as another widely applied method to take into account solvent effects in computational chemistry [21]. A schematic illustration of implicitly defined solvents for boric acid and borate is demonstrated in figure 4.1.

Computation of k_{3-4} in pure water has already been reported by Liu and Tossel [14]. However, the IEF-PCM and CPCM continuum solvation models they employed at the HF/6-31G* level of theory for this purpose are commonly found to be less accurate compared to more recent solvation models like the SMx family of methods and higher levels of theory, as we have shown in a recent study [21]. Therefore, in addition to IEF-PCM and CPCM, we also study the SMD solvation model for a high level of theory for pure and saline water.
We also employ the implicit solvent approach for verifying the suitability of our selected cluster sizes for appropriately taking into account the long-range interactions for the NMA with explicit solvent. To that end, in addition to 60 clusters studied in vacuo, we also studied

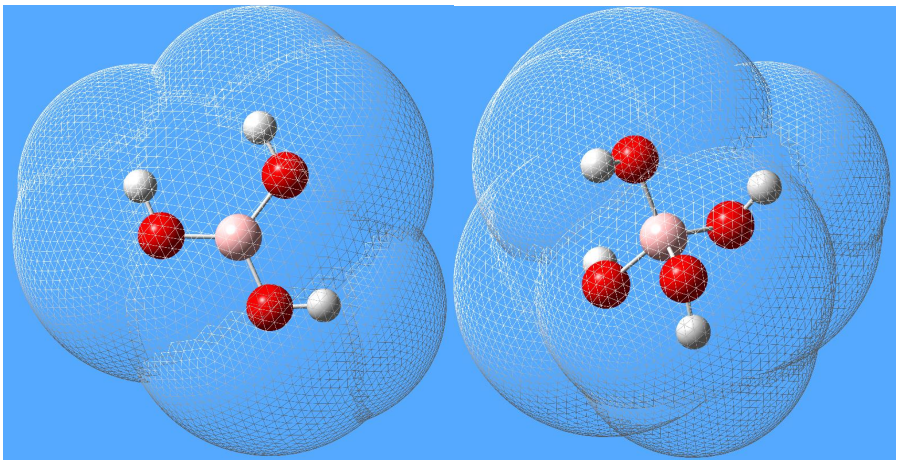


Fig. 4.1 Illustration of implicitly defined solvents in boric acid and borate.

20 clusters placed inside a cavity of implicitly defined solvents as implemented in several works [33-37].

4.2.5 Theoretical evaluation of equilibrium constant via path integral molecular dynamics

In addition to NMA which is the classic theoretical approach for evaluation of equilibrium constant of isotope exchange reactions, another rigorous approach that has recently been widely used for this purpose is the thermodynamic integration of kinetic energies associated with isotope effects via Path Integral Molecular Dynamics (PIMD) [17,38-40]. Based on this approach, the equilibrium constant of an isotope exchange reaction is calculated as:

$$K_{eq} = -\frac{3}{2}\ln\left(\frac{m_2}{m_1}\right) + \frac{1}{k_B T}\int_{m_1}^{m_2} \frac{\langle K_E(m)\rangle}{m} dm, \tag{4.15}$$

where m_i is the mass of isotope i and K_E is the kinetic energy. To calculate the equilibrium constant via Eq. (4.15), the ensemble average of kinetic energies for a number of masses alchemically transformed between $m_1$and m_2 should be computed via PIMD to allow calculating the area under the $\frac{\langle K_E(m)\rangle}{m}$ versus m curve, as required by Eq. (4.15). Although alchemically transforming the isotope masses can be done arbitrarily, doing this via the following switching function:

$$m(\lambda) = \frac{m_1 m_2}{\left(\lambda\sqrt{m_1} + (1-\lambda)\sqrt{m_2}\right)^2}, \tag{4.16}$$

where λ changes from 0 to 1 to yield transformed masses between m_1 and m_2, has found to be more advantageous [41] and will be employed in the present study as well. One common way to evaluate the kinetic energies required by Eq. (4.15) via PIMD is based on the primitive kinetic energy estimator defined as:

$$\langle K_E(m)\rangle^{(\text{primitive})} = \frac{3Pk_BT}{2} - \left\langle \sum_{i=1}^{P} \frac{1}{2} m\omega_p^2 (r_i - r_{i-1})^2 \right\rangle, \tag{4.17}$$

in which P is the number of beads, $\omega_p = \sqrt{Pk_BT/\hbar}$ is the ring polymer frequency and r_i is the position of the ith bead of the studied isotope. However, due to the well-known drawback of the primitive estimator of kinetic energy which is its increasing variance for increasing the number of beads [42], the virial estimator of kinetic energy defined as:

$$\langle K_E(m)\rangle^{(\text{virial})} = \frac{3k_BT}{2} + \left\langle \frac{1}{2P} \sum_{i=1}^{P} (r_i - r_c) \frac{\partial U}{\partial r_i} \right\rangle, \tag{4.18}$$

is more commonly used for this purpose. Here, r_c is the position of the centroid of the beads, and $\frac{\partial U}{\partial r_i}$ is the net force on the ith bead due to the interaction of the atoms in the same replica and without the inter-bead interactions included.

In the present study, in the evaluation of kinetic energy, we exploit the features of both abovementioned estimators to improve numerical approximation of the integral in Eq. (4.15). Accordingly, considering that the primitive estimator implies a linear dependency between the kinetic energies and switched masses, we fit a linear curve of the form $\langle K_E(m)\rangle = a + bm$, where a and b are constants and the kinetic energies are calculated based on the virial estimator. It allows us to reduce the numerical noises of each mass using all kinetic energy data for all other masses and also be able to solve the integral in Eq. (4.15) analytically.

Compared to NMA, thermodynamic integration through PIMD offers a number of advantages such as taking into account the anharmonicity effects as well as an appropriate sampling of canonical ensemble and the possibility of studying multi-component solvents. As a result, we employed this method for estimating k_{3-4} for the boron isotope fractionation in both pure and saline water.

4.3 Computational details

4.3.1 Evaluation of k_{3-4} via NMA

To calculate k_{3-4} based on NMA with explicit solvent, 60 randomly selected clusters of 64 water molecules with one boric acid or borate ion in the center of the cluster were extracted

from a trajectory of configurations produced by Molecular Dynamics (MD) simulations. These MD simulations were carried out for a simulation box containing 1200 water and one boric acid or borate molecule under NVT conditions. The box size was fine-tuned to yield a density equal to 0.995 g/mol which was close to the density of water at standard conditions. The inter-atomic interactions in the MD simulations were evaluated using the GFN2-xTB method [43] and the temperature was controlled using a Nose-Hoover Chain thermostat with three chains and 50 fs relaxation time. We ran the MD simulations in CP2K [50] for 10 ns after 1ns equilibration with 0.5 fs time step and took snapshots every 100ps to generate the required configurations via screening the solute and 64 water molecules closest to it.

For the generated configurations, the normal modes were computed based on the partial normal mode analysis explained in section 4.2.2. To that end, the geometries of the solutes were relaxed while the solvent geometries were frozen. Considering that the employed level of theory plays a significant role in the accuracy of evaluated k_{3-4} as discussed earlier, we employed the DSD-PBEP86/QZVPP level of theory for optimization and calculations of normal modes of the solute as a rigorous and accurate method for evaluating thermodynamics quantities [19,20] and scaled the vibrational frequencies by the recommended scaling factor of 0.9971 for improving the accuracy of this method [44]. However, for the solvent molecules, for which the normal mode calculations are not required and thus employing a high level of theory loses its importance, we used the B3LYP/6-311++G(2d,p) level of theory and computed the total energy of the cluster via the ONIOM approach.

Although clusters with 64 water molecules seem to be large enough to appropriately solvate the studied solutes, as a further verification for 20 of the studied configurations we also computed the normal modes of those clusters placed in the cavity of a continuously defined solvent based on the IEF-PCM continuum solvation model.

For the computation of k_{3-4} solely with the implicit solvent approach, we calculated the normal modes using the DSD-PBEP86/QZVPP level of theory and the SMD, CPCM, and IEF-PCM continuum solvation models for both pure and saline water. For the CPCM solvation model, in addition to the original implementation of this model in Gaussian 16, we also studied scaling the dielectric constant of the solvent via:

$$\widetilde{\varepsilon}(\varepsilon, x) = \frac{\varepsilon + x}{x + 1} \tag{4.19}$$

for $x = 0.5$, as this was found to be more efficient in improving the accuracy of the CPCM method [21,45].

For calculation of k_{3-4} in saline water, the dielectric constant of the saline water required by the continuum solvation models with the value of 70.35 recommended by Lang et al. [46] was employed. All the computations were carried out using the Gaussian 16 software [47].

4.3.2 Evaluation of k_{3-4} via PIMD

To study the anharmonic effects as well as the contributions from the configurational entropies, we employed PIMD based on the guidelines provided in section 4.2.5. To that end, we studied PIMD with 4, 8, 16, 32, and 64 beads. Using switching constants λ linearly distributed between zero and one, 9 isotopic masses of boron varying between 10.01294 to 11.009305 were assigned via the switching function defined in Eq. (4.16). The kinetic energies evaluated based on the virial estimator for the studied masses were then used to calculate k_{3-4} as described in section 4.2.5.

The PIMD simulations were carried out in CP2K within the staging approach [48] for a duration of 500 ps, 0.5 fs time step, temperature controlled via Nose-Hoover Chain thermostat with three chains, periodic boundary condition, and the interactions evaluated via GFN2-xTB method [43].

For both pure and saline water, the PIMD simulations were carried out in canonical ensemble and the box size was set to yield 0.995 gr/mol density for the system. For the pure water, the simulation box contained 64 water molecules and one boric acid or borate. However, for the saline water, to set a reasonable molar ratio for water and NaCl as recommended by Zeron et al. [49], we constructed a simulation box containing one boric acid or borate, 5 NaCl, and 572 water molecules.

4.4 Results and discussion

The RPFs of boric acid and borate isotopomers calculated for 60 configurations based on the NMA with explicit solvent and the energy of the solute in each one evaluated by the IEF-PCM solvation model are reported in table 4.2. The calculated k_{3-4} values using different integrations of these data are reported in table 4.3. The results obtained via NMA with explicit solvent approach and Boltzmann weighted integration show the best agreement with the upper limit of the extrapolated results reported by Rustad et al. (k_{3-4}=1.028) for the same approach and the highest level of theory they employed [16]. Nevertheless, all these results are slightly lower than the experimental value of 1.0308±0.0023 reported by Klochko et al. [15].

Note that although we report integrated results for 60 configurations, the arithmetic averaging and the Boltzmann-weighted averaging result in k_{3-4} values of 1.0263 and 1.0279, respectively. These two numbers still show a deviation similar to the one between the upper and lower limits of the extrapolated results reported by Rustad et al. [16]. However, the results obtained via the Boltzmann weighted averaging are in much better agreement with those we obtained via NMA and the implicit solvent approach as well as PIMD reported

in the following, and are almost within the uncertainty of the experimentally determined values of 1.0308±0.0023 reported by Klochko et al. [15]. This confirms the robustness of the Boltzmann weighted averaging.

As discussed earlier, for 20 of the studied configurations, we also repeated the optimization and NMA for the same clusters placed in the cavity of an implicitly defined solvent. However, our results showed almost exactly the same values up to 3 significant figures between the two approaches which verifies that the explicit cluster size is large enough to appropriately solvate the studied solutes.

The evaluated RPFs and k_{3-4} calculated via NMA with implicit solvent are reported in table 4.4. These results show an excellent agreement with those obtained via PIMD. These results also imply the possibility of obtaining accurate results for the evaluation of isotope fractionation via the implicit solvent approach despite its much lower computational challenge and cost. However, it should be noted that all this has become possible via carefully parameterizing the employed continuum solvation models to reproduce solvation free energies of numerous solutes in water and thus empirically taking into account and correcting limitations due to anharmonicity or inefficient canonical sampling [21]. The advantage of NMA with explicit solvent is that it yields results that are obtained entirely from the first principles and therefore can be more reliably employed for new solvents or solutes.

According to the results of NMA with implicit solvent, the estimated k_{3-4} for pure and saline water are almost the same, similar to the results we obtained via PIMD discussed in the following. These results are also in contrast to the experimentally determined values reported by Klochko et al. [15] which suggested more diverse values for k_{3-4} in pure and saline water equal to 1.0308±0.0023 and 1.0272±0.0003, respectively. However, considering that the uncertainty in the experimentally measured data reported by Klochko et al. for k_{3-4} in pure water is much higher than those observed for saline water, implying some overlapping values in between, we can conclude that k_{3-4} is almost the same in pure and saline water.
The evaluated k_{3-4} obtained via PIMD for pure and saline water for the different path integral bead numbers are reported in table 4.5. These results show that ring polymers with 32 beads yield converged results for estimated k_{3-4}. These results also show a good agreement with the theoretically predicted results obtained via NMA as well as the experimentally measured values. Similar to the results obtained via NMA with implicit solvent, here also the results imply that the k_{3-4} for both pure and saline water are not significantly different.

Table 4.2 The RPFs and conformer energies (ε_i, Hartree/molecule) for boric acid and borate.

	RPF-$B(OH)_3$	ε_i-$B(OH)_3$	RPF-$B(OH)_4^-$	ε_i- $B(OH)_4^-$		RPF-$B(OH)_3$	ε_i-$B(OH)_3$	RPF-$B(OH)_4^-$	ε_i-$B(OH)_4^-$
1	1.232289	-252.260	1.201585	-328.159	**31**	1.233663	-252.263	1.200363	-328.162
2	1.233101	-252.260	1.200963	-328.161	**32**	1.231332	-252.260	1.200272	-328.160
3	1.233308	-252.262	1.202624	-328.159	**33**	1.231903	-252.261	1.199513	-328.160
4	1.229193	-252.258	1.200637	-328.160	**34**	1.233421	-252.258	1.199002	-328.161
5	1.232274	-252.263	1.201156	-328.157	**35**	1.231168	-252.257	1.200133	-328.159
6	1.229627	-252.257	1.201668	-328.160	**36**	1.232375	-252.257	1.197710	-328.160
7	1.234888	-252.263	1.198716	-328.161	**37**	1.231969	-252.264	1.201591	-328.159
8	1.232739	-252.261	1.203835	-328.157	**38**	1.231112	-252.261	1.202307	-328.160
9	1.233192	-252.260	1.200803	-328.161	**39**	1.233183	-252.261	1.198171	-328.159
10	1.231513	-252.261	1.200002	-328.161	**40**	1.233118	-252.262	1.198609	-328.161
11	1.230923	-252.258	1.197027	-328.159	**41**	1.233520	-252.263	1.201498	-328.158
12	1.228097	-252.257	1.196828	-328.158	**42**	1.231855	-252.261	1.197847	-328.161
13	1.232589	-252.261	1.195636	-328.160	**43**	1.231172	-252.261	1.204232	-328.160
14	1.230915	-252.259	1.198634	-328.160	**44**	1.229711	-252.259	1.198582	-328.159
15	1.231508	-252.259	1.200120	-328.158	**45**	1.232624	-252.261	1.198204	-328.162
16	1.228483	-252.261	1.201247	-328.161	**46**	1.233064	-252.263	1.201458	-328.160
17	1.234450	-252.262	1.203933	-328.159	**47**	1.230490	-252.259	1.202082	-328.159
18	1.229549	-252.256	1.201923	-328.160	**48**	1.231707	-252.262	1.203092	-328.160
19	1.230690	-252.256	1.198874	-328.158	**49**	1.232594	-252.262	1.199778	-328.161
20	1.229597	-252.257	1.196579	-328.160	**50**	1.230889	-252.260	1.201935	-328.160
21	1.231391	-252.260	1.196280	-328.160	**51**	1.230689	-252.257	1.202002	-328.158
22	1.232017	-252.261	1.198901	-328.161	**52**	1.234509	-252.261	1.196724	-328.159
23	1.234101	-252.263	1.200880	-328.161	**53**	1.230561	-252.258	1.202011	-328.160
24	1.229961	-252.258	1.204097	-328.160	**54**	1.232139	-252.260	1.200349	-328.159
25	1.231632	-252.258	1.204082	-328.159	**55**	1.233125	-252.263	1.201442	-328.157
26	1.234181	-252.261	1.198961	-328.158	**56**	1.231232	-252.257	1.203445	-328.159
27	1.231185	-252.259	1.200437	-328.159	**57**	1.229698	-252.254	1.202555	-328.160
28	1.230308	-252.260	1.199040	-328.162	**58**	1.233923	-252.261	1.201612	-328.162
29	1.231356	-252.259	1.195705	-328.160	**59**	1.230443	-252.258	1.197761	-328.160
30	1.231660	-252.260	1.200217	-328.161	**60**	1.229274	-252.255	1.195276	-328.159

Table 4.3 Integrated results of RPFs and the evaluated k_{3-4}.

	$B(OH)_3$	$B(OH)_4^-$	k_{3-4}
Arithmetic averaging	1.2317	1.2002	1.0263
Boltzmann weighted averaging	1.2330	1.1997	1.0278

Table 4.4 Calculated RPFs and the evaluated k_{3-4} via different continuum solvation models for pure and saline water.

	Pure water			Saline water		
	$B(OH)_3$	$B(OH)_4^-$	k_{3-4}	$B(OH)_3$	$B(OH)_4^-$	k_{3-4}
SMD	1.2206	1.1820	1.0327	1.2207	1.1820	1.0327
IEF-PCM	1.2267	1.1902	1.0306	1.2267	1.1902	1.0306
CPCM	1.2266	1.1901	1.0307	1.2266	1.1901	1.0306
CPCM(x=0.5)	1.2267	1.1902	1.0307	1.2267	1.1902	1.0307

Table 4.5 Evaluated k_{3-4} via PIMD for pure and saline water.

Nr. beads	Pure water	Saline water
4	1.013377	1.013113
8	1.023831	1.02616
16	1.031404	1.030177
32	1.033596	1.030278
64	1.030522	1.030118

4.5 Conclusion

In the present study, theoretical evaluation of k_{3-4} using high-level NMA with the explicit solvent approach, three different implicit solvent approaches, and PIMD was studied. By comparing various high-precision theoretical methods employed in the present study with those reported by Klochko et al. [15] via experimental techniques and theoretically reported results of Rustad et al. [16], we can propose the most probable estimation of k_{3-4} within 1.028 to 1.030 for both pure and saline water.

Acknowledgment

This project has received funding in the framework of European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 643084. The authors wish to thank Simone Knief and Karsten Balzer in the computing center of Christian-Albrechts University of Kiel for their assistance and technical supports for running the computations on the high-performance computation hardware of Kiel University.

References:

1. Dickson, A. G.; Goyet, C. Handbook of methods for the analysis of the various parameters of the carbon dioxide system in sea water. Version 2; Oak Ridge National Lab., TN (United States): 1994.
2. Zeebe, R. E.; Sanyal, A.; Ortiz, J. D.; Wolf-Gladrow, D. A., A theoretical study of the kinetics of the boric acid–borate equilibrium in seawater. Marine Chemistry 2001, 73 (2), 113-124.
3. Vengosh, A.; Kolodny, Y.; Starinsky, A.; Chivas, A. R.; McCulloch, M. T., Coprecipitation and isotopic fractionation of boron in modern biogenic carbonates. Geochimica et Cosmochimica Acta 1991, 55 (10), 2901-2910.
4. Hemming, N. G.; Hanson, G. N., Boron isotopic composition and concentration in modern marine carbonates. Geochimica et Cosmochimica Acta 1992, 56 (1), 537-543.
5. Pearson, P. N.; Palmer, M. R., Atmospheric carbon dioxide concentrations over the past 60 million years. Nature 2000, 406 (6797), 695-699.
6. Tyrrell, T.; Zeebe, R. E., History of carbonate ion concentration over the last 100 million years. Geochimica et Cosmochimica Acta 2004, 68 (17), 3521-3530.
7. De La Vega, E.; Foster, G. L.; Martínez-Botí, M. A.; Anagnostou, E.; Field, M. P.; Kim, M. H.; Watson, P.; Wilson, P. A., Automation of boron chromatographic purification for d11B analysis of coral aragonite. Rapid Communications in Mass Spectrometry 2020, 34 (11), e8762.
8. Carrey, R.; Ballesté, E.; Blanch, A. R.; Lucena, F.; Pons, P.; López, J. M.; Rull, M.; Solà, J.; Micola, N.; Fraile, J., Combining multi-isotopic and molecular source tracking methods to identify nitrate pollution sources in surface and groundwater. Water Research 2021, 188, 116537.
9. Noireaux, J.; Sullivan, P. L.; Gaillardet, J.; Louvat, P.; Steinhoefel, G.; Brantley, S. L., Developing boron isotopes to elucidate shale weathering in the critical zone. Chemical Geology 2021, 559, 119900.
10. Kakihana, H.; Kotaka, M.; Satoh, S.; Nomura, M.; Okamoto, M., Fundamental studies on the ion-exchange separation of boron isotopes. Bulletin of the Chemical Society of Japan 1977, 50 (1), 158-163.
11. Oi, T.; Yanase, S., Calculations of reduced partition function ratios of hydrated monoborate anion by the ab initio moleculaSr orbital theory. Journal of nuclear science and technology 2001, 38 (6), 429-432.
12. Yamahira, M.; Oi, T., Calculations of reduced partition function ratios of hydrated boric acid molecule by the ab initio molecular orbital theory. Journal of nuclear science and technology 2004, 41 (8), 832-836.
13. Zeebe, R. E., Stable boron isotope fractionation between dissolved B(OH)3 and B(OH)4. Geochimica et Cosmochimica Acta 2005, 69 (11), 2753-2766.
14. Liu, Y.; Tossell, J. A., Ab initio molecular orbital calculations for boron isotope fractionations on boric acids and borates. Geochimica et Cosmochimica Acta 2005, 69 (16), 3995-4006.

15. Klochko, K.; Kaufman, A. J.; Yao, W.; Byrne, R. H.; Tossell, J. A., Experimental measurement of boron isotope fractionation in seawater. Earth and Planetary Science Letters 2006, 248 (1-2), 276-285.
16. Rustad, J. R.; Bylaska, E. J.; Jackson, V. E.; Dixon, D. A., Calculation of boron-isotope fractionation between B (OH) 3 (aq) and B (OH) 4-(aq). Geochimica et Cosmochimica Acta 2010, 74 (10), 2843-2850.
17. Dupuis, R.; Benoit, M.; Tuckerman, M. E.; Meheut, M., Importance of a fully anharmonic treatment of equilibrium isotope fractionation properties of dissolved ionic species as evidenced by Li+ (aq). Accounts of chemical research 2017, 50 (7), 1597-1605.
18. Rustad, J. R.; Bylaska, E. J., Ab initio calculation of isotopic fractionation in B (OH) 3 (aq) and BOH4-(aq). Journal of the American Chemical Society 2007, 129 (8), 2222-2223.
19. Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S., A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. Physical Chemistry Chemical Physics 2017, 19 (48), 32184-32215.
20. Kesharwani, M. K.; Brauer, B.; Martin, J. M., Frequency and zero-point vibrational energy scale factors for double-hybrid density functionals (and other selected methods): can anharmonic force fields be avoided? The Journal of Physical Chemistry A 2014, 119 (9), 1701-1714.
21. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. Nature Communications 2021,12(1),1-7.
22. McQuarrie, D. A., Statistical thermodynamics. 1973.
23. Bigeleisen, J.; Mayer, M. G., Calculation of equilibrium constants for isotopic exchange reactions. The Journal of Chemical Physics 1947, 15 (5), 261-267.
24. Sato, A.; Bernier-Latmani, R.; Hada, M.; Abe, M., Ab initio and steady-state models for uranium isotope fractionation in multi-step biotic and abiotic reduction. Geochimica et Cosmochimica Acta 2021.
25. Pinilla, C.; De Moya, A.; Rabin, S.; Morard, G.; Roskosz, M.; Blanchard, M., First-principles investigation of equilibrium iron isotope fractionation in Fe1-xSx alloys at Earth's core formation conditions. Earth and Planetary Science Letters 2021, 569, 117059.
26. Balan, E.; Noireaux, J.; Mavromatis, V.; Saldi, G. D.; Montouillout, V.; Blanchard, M.; Pietrucci, F.; Gervais, C.; Rustad, J. R.; Schott, J., Theoretical isotopic fractionation between structural boron in carbonates and aqueous boric acid and borate ion. Geochimica et Cosmochimica Acta 2018, 222, 117-129.
27. Pinilla, C.; Blanchard, M.; Balan, E.; Natarajan, S. K.; Vuilleumier, R.; Mauri, F., Equilibrium magnesium isotope fractionation between aqueous Mg2+ and carbonate minerals: insights from path integral molecular dynamics. Geochimica et Cosmochimica Acta 2015, 163, 126-139.
28. Feng, C.; Qin, T.; Huang, S.; Wu, Z.; Huang, F., First-principles investigations of equilibrium calcium isotope fractionation between clinopyroxene and Ca-doped orthopyroxene. Geochimica et Cosmochimica Acta 2014, 143, 132-142.
29. Huang, F.; Chen, L.; Wu, Z.; Wang, W., First-principles calculations of equilibrium Mg isotope fractionations between garnet, clinopyroxene, orthopyroxene, and olivine: implications for Mg isotope thermometry. Earth and Planetary Science Letters 2013, 367, 61-70.
30. Javoy, M.; Balan, E.; Méheut, M.; Blanchard, M.; Lazzeri, M., First-principles investigation of equilibrium isotopic fractionation of O-and Si-isotopes between refractory solids and gases in the solar nebula. Earth and Planetary Science Letters 2012, 319, 118-127.

31. Kowalski, P. M.; Jahn, S., Prediction of equilibrium Li isotope fractionation between minerals and aqueous solutions at high P and T: an efficient ab initio approach. Geochimica et Cosmochimica Acta 2011, 75 (20), 6112-6123.
32. Tuckerman, M., Statistical mechanics: theory and molecular simulation. Oxford university press: 2010.
33. da Silva, E. F.; Svendsen, H. F.; Merz, K. M., Explicitly representing the solvation shell in continuum solvent calculations. The Journal of Physical Chemistry A 2009, 113 (22), 6404-6409.
34. Hayes, J. M.; Bachrach, S. M., Effect of micro and bulk solvation on the mechanism of nucleophilic substitution at sulfur in disulfides. The Journal of Physical Chemistry A 2003, 107 (39), 7952-7961.
35. da Silva, C. O.; Mennucci, B.; Vreven, T., Combining microsolvation and polarizable continuum studies: New insights in the rotation mechanism of amides in water. The Journal of Physical Chemistry A 2003, 107 (34), 6630-6637.
36. Fernández-Ramos, A.; Miller, J. A.; Klippenstein, S. J.; Truhlar, D. G., Modeling the kinetics of bimolecular reactions. Chemical reviews 2006, 106 (11), 4518-4584.
37. Li, Y.; Hartke, B., Assessing Solvation Effects on Chemical Reactions with Globally Optimized Solvent Clusters. ChemPhysChem 2013, 14 (12), 2678-2686.
38. Wang, L.; Ceriotti, M.; Markland, T. E., Quantum kinetic energy and isotope fractionation in aqueous ionic solutions. Physical Chemistry Chemical Physics 2020, 22 (19), 10490-10499.
39. Daengngern, R.; Kobayashi, O.; Kungwan, N.; Ngaojampa, C.; Tachikawa, M., Nuclear quantum and H/D isotope effects on three-centered bonding diborane: Path integral molecular dynamics simulations. International Journal of Quantum Chemistry 2020, 120 (10), e26179.
40. Brela, M. Z.; Prah, A.; Boczar, M.; Stare, J.; Mavri, J., Path integral calculation of the hydrogen/deuterium kinetic isotope effect in monoamine oxidase a-catalyzed decomposition of benzylamine. Molecules 2019, 24 (23), 4359.
41. Marsalek, O.; Chen, P.-Y.; Dupuis, R.; Benoit, M.; Meheut, M.; Bacic, Z.; Tuckerman, M. E., Efficient calculation of free energy differences associated with isotopic substitution using path integral molecular dynamics. Journal of chemical theory and computation 2014, 10 (4), 1440-1453.
42. Herman, M.; Bruskin, E.; Berne, B., On path integral Monte Carlo simulations. The Journal of Chemical Physics 1982, 76 (10), 5150-5155.
43. Bannwarth, C.; Ehlert, S.; Grimme, S., GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. Journal of chemical theory and computation 2019, 15 (3), 1652-1671.
44. Kesharwani, M. K.; Sylvetsky, N.; Martin, J. M. In Surprising performance for vibrational frequencies of the distinguishable clusters with singles and doubles (DCSD) and MP2. 5 approximations, AIP Conference Proceedings, AIP Publishing: 2017; p 030005.
45. Klamt, A.; Moya, C.; Palomar, J., A comprehensive comparison of the IEFPCM and SS (V) PE continuum solvation methods with the COSMO approach. Journal of chemical theory and computation 2015, 11 (9), 4220-4225.
46. Lang, R.; Zhou, Y.; Utku, C.; Le Vine, D., Accurate measurements of the dielectric constant of seawater at L band. Radio Science 2016, 51 (1), 2-24.
47. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., Gaussian 16. Gaussian, Inc. Wallingford, CT: 2016.

48. Tuckerman, M.; Berne, B., Vibrational relaxation in simple fluids: Comparison of theory and simulation. The Journal of chemical physics 1993, 98 (9), 7301-7318.
49. Zeron, I. M.; Gonzalez, M. A.; Errani, E.; Vega, C.; Abascal, J. L., "In Silico" Seawater. Journal of Chemical Theory and Computation 2021, 17 (3), 1715-1725.
50. Hutter, J., Iannuzzi, M., Schiffmann, F. & VandeVondele, J. cp2k: atomistic simulations of condensed matter systems. Wiley Interdisciplinary Reviews: Computational Molecular Science 4, 15-25 (2014).

Chapter 5

Thermodynamically effective molecular surfaces

Project overview and motivation:

Molecular surfaces are among the parameters which have the highest impact on the thermodynamics of solution and are required in a large number of cutting-edge scientific and technological applications. The scientific works which are proposed in the last century to describe the connection between molecular surfaces and condensed phase thermodynamics are mainly derived empirically and therefore, suffer from accuracy and domain of applications. Furthermore, the evaluation of molecular surfaces in conventional methods mainly involves empirical corrections and single-point measurements and therefore, totally overlooks the thermodynamic efficiency of those surfaces. It motivated us to study the theoretical dependency between molecular surfaces and condensed phase thermodynamics and use it to define alternative molecular surfaces which are only slightly different than the conventionally accepted ones but are significantly more efficient in studying the thermodynamics of condensed phase.

Novelty aspects:

- Proposing thermodynamically effective molecular surfaces and demonstrating their significant efficiency in evaluating thermodynamic quantities in solution compared to conventionally defined molecular surfaces
- The theoretically proposed relationship for describing temperature dependence of vaporization enthalpy which is currently the only theoretically derived correlation yielding accurate results for the wide temperature ranges

- The proposed theoretically obtained connection between solution thermodynamics and molecular surfaces which invalidates the two other empirically proposed models with a history of almost a century

Connection to other chapters:

The thermodynamically effective molecular surfaces are in close connection to the cavity surfaces based on the implicit solvent approaches studied in chapters 4, 8, and 9. The machine learning models and correlations which are proposed in chapter 7 for a more practical evaluation of vaporization enthalpy at various temperatures is entirely based on the findings and achievements presented in this project.

Contributions:

Conceiving the idea of thermodynamically effective molecular surfaces, major contribution in method development, carrying out all the computations, major contribution in writing the manuscript.

Publication status:

This work is in the submission process and its preprint is available online (DOI:10.21203/rs.3.rs-816803/v1).

Thermodynamically effective molecular surfaces for more efficient study of condensed-phase thermodynamics

Amin Alibakhshi[1,*], Bernd Hartke[1]
Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, 24118 Kiel, Germany
Corresponding author: alibakhshi@pctc.uni-kiel.de

Abstract

Evaluation of molecular surfaces is of key importance in a wide range of cutting-edge scientific fields and technologies. Due to its extensive applications, numerous methods such as van-der-Waals and solvent-accessible surface areas, have been proposed to provide an estimation of molecular surfaces. Each one of these methods and various parameterizations proposed for them, typically yield quite diverse estimations of molecular surfaces which necessitates employing ad-hoc modifications in their respective applications to accommodate inappropriately defined molecular surfaces. The other main shortcoming of the currently defined molecular surfaces stems from their parameterization based on single-point physical quantities, though their main application is typically the evaluation of thermodynamic quantities in the condensed phase.
To address all these limitations, in the present study we propose *thermodynamically effective* molecular surfaces which, unlike the previously proposed surfaces, can be defined only uniquely, can be measured experimentally for each molecule directly and straightforwardly, and is defined based on a well-characterized theoretically described dependency between molecular surfaces and solution thermodynamics. Although the thermodynamically effective surfaces are defined through a novel thermodynamics approach, we show that they are very close to van-der-Waals surfaces. Nevertheless, by extensively benchmarking we demonstrate the superior efficiency of the newly defined surfaces in substantially improving the predictability of multiple thermodynamics quantities in solution without requiring any ad-hoc modification, compared to the conventionally accepted ones.

5.1 Introduction

The surface area of molecules is one of the key parameters influencing the thermodynamics of the condensed phase. Evaluation of molecular surfaces plays a key role in a large number of cutting-edge scientific fields and technologies such as drug discovery [1,2], catalysis

[3], molecular biology [4-6], molecular genetics [7], and nanotechnology [8,9]. One of the main applications of molecular surface estimation is theoretical evaluation of solution thermodynamics via continuum solvation models [10] which is an extensively applied method in very diverse scientific fields, ranging from catalysis [11,12], advanced nanomaterials [13], surface science [14], or mechanisms of chemical reactions in the condensed phase [15,16] to unraveling the activity mechanism of coronavirus [17].

Employing molecular surfaces for unraveling scientific challenges in the condensed phase has been an active scientific area with a history of more than a century. One of the earliest attempts in this regard dates back to 1886 and was proposed by Eötvös. He suggested a proportionality between the free energy per unit of interfacial surface, i.e. the surface tension, and the surface area of liquid-phase molecules [18]. This work was indeed one of the earliest examples of experimental evaluation of molecular surfaces, which was achieved by assuming solution-phase molecules as perfect spheres, allowing to evaluate molecular surface area via liquid molar volume. Although the assumption of perfect spheres is the simplest approach to get a rough estimation of molecular surfaces, it satisfactorily holds for mono-atomic molecules. Accordingly, the earliest successful applications of molecular surfaces to study solution thermodynamics both exploit surfaces determined via perfect sphere assumption and are mainly limited to noble gases [19-22].

In 1964, the van-der-Waals (vdW) surface area concept was proposed in the pioneering work of Bondi [23], which became the cornerstone of more advanced molecular surfaces such as Solvent Excluded Surfaces (SES) and Solvent Accessible Surface Area (SASA). Since then, a large number of methods and algorithms such as multiple variants of solvent excluded or solvent accessible surface areas have been proposed to provide evaluations of molecular surfaces in solution. This wide variety of methods typically yields quite diverse estimations of molecular surfaces, as can be seen in a comparison of vdW and SAS molecular surfaces of ethylene depicted in figure 5.1. For a broader comparison, we provide computed molecular surfaces for 215 molecules via different parameterizations of the vdW method as supplementary material. These data imply that selecting the most appropriate method is not a trivial task.

Surprisingly, despite the diversity of methods and molecular surface approximations they yield, there are numerous examples of reporting successful applications for each one of these methods. This is mainly because the majority of these research works are based on empirically defined relationships between solution thermodynamics and molecular surfaces [24,25]. Inaccuracies due to deviations of the employed molecular surfaces from the actual values are then corrected via ad-hoc modifications and parametrizations, mainly applied to atomic radii. For example, while the Gaussian 03 software package used SES surfaces and

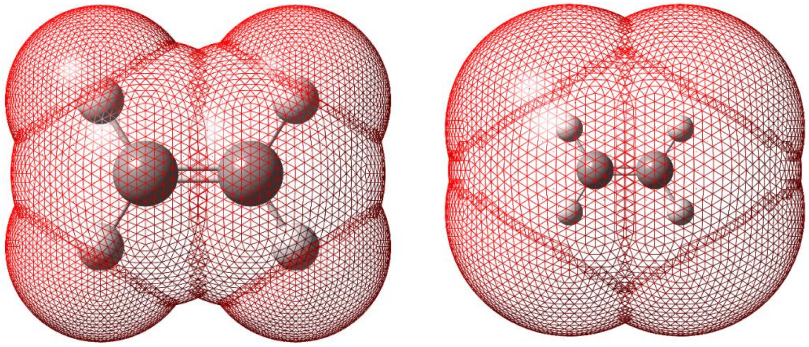


Fig. 5.1 Comparison of vdW (left) and SAS (right) surfaces in ethylene.

UA0 atomic radii as default for computing solvent effects based on the polarizable continuum solvation models, for the latest release of the same software, vdW surfaces and UFF atomic radii are considered as default [26]. Similarly, the most widely applied continuum solvation models exploit their specifically defined molecular surfaces and parameterizations of atomic radii [10].

One main reason behind this diversity in defining and parameterizing molecular surfaces and requiring such ad-hoc modification is that although the main application of molecular surfaces is commonly for studying solution-phase thermodynamics, they are typically parameterized for reproducing other target quantities. For example, the Bondi parameterization of atomic radii has been done using physical quantities like X-ray diffraction data, gas-phase kinetic collision cross-section, and liquid density as target quantities [23] while the UFF or UA0 atomic radii are parametrized against bond distances [27].
It indeed stems from the unavailability of a rigorous theoretical method that allows precise and analytical characterization of the relationship between molecular surfaces and solution thermodynamics without ad-hoc modifications or parameterization.

Obviously, such a theoretical method offers a number of advantages. First and foremost, it allows defining molecular surfaces with physical significance, without requiring ad-hoc modifications, and thus uniquely definable. Furthermore, it provides the possibility of evaluating the performance of various methods and parametrizations in reproducing those reference molecular surfaces. Last but not least, it allows for better understanding and treating some of the main challenges in theoretical studies of solvation, such as the appropriate treatment of solvent effects in continuum solvation models.

To achieve this goal, in the present study we exploit a theoretically derived relationship describing temperature dependence of vaporization enthalpy to the molecular surfaces in solution which is an extension of a recent study [28] leading to a remarkable improvement of the formerly reported results. Among all potential thermodynamics quantities of solution

which can be analytically related to molecular surfaces for this purpose, the vaporization enthalpy, which is employed in the present study possesses a number of obvious advantages. The main one is that vaporization enthalpy can be directly determined experimentally, while free energy or entropy can only be determined indirectly and via measuring the temperature dependence of vaporization enthalpy or equilibrium vapor pressures at multiple temperatures, which implies the accumulation of errors inherent in both experimental measurements and the subsequent computations. The more convenient experimental procedure for enthalpy measurement has also made accurate benchmark datasets more readily available, which is another advantage of using vaporization enthalpy. Finally, evaluation of molecular surfaces via vaporization enthalpy is not only both more accurate and less challenging but also once it is found, it can be conveniently used to obtain other thermodynamic quantities, via the fundamental thermodynamics relationships, as shown in section 5.4.3.

5.2 Theory

By considering vaporization as a dynamic process at which evaporation and condensation have the same rates and equating the rates of evaporation and condensation described by transition state theory and some manipulations, the ratio of partition functions of the gas and liquid phases is obtained as [28]:

$$\frac{Q^g}{Q^s} = \frac{N_a\, P^{sat}}{P}\left(\frac{k_B T}{2\pi m}\right)^{1/2} \frac{h}{k_B T[n_s]} \exp\left(\frac{\Delta\varepsilon_{sg}}{k_B T}\right), \tag{5.1}$$

where N_a is Avogadro's constant, P^{sat} is the saturation vapor pressure of the liquid, $\Delta\varepsilon_{sg}$ is the energy for moving one molecule from the liquid surface to the gas phase, and k_B and h are Boltzmann and Planck constants, respectively. Using the statistical thermodynamics relationship between the energy and partition function stated as:

$$\langle\varepsilon\rangle = k_B T^2 \frac{\partial \ln(Q)}{\partial T}, \tag{5.2}$$

and with some algebraic manipulations, it can be shown that the temperature dependence of the vaporization enthalpy follows [28]:

$$\Delta h_{vap} = \Delta\varepsilon_{bs} - \frac{k_B}{2} T\ln(T) - T\int \frac{\Delta\varepsilon_{bs}}{T^2} dT + CT, \tag{5.3}$$

where C is a constant and $\Delta\varepsilon_{bs}$ is the energy required for moving one molecule from the bulk of the liquid to the surface. Evaluation of $\Delta\varepsilon_{bs}$ via experimentally measurable quantities can be achieved using the fundamental thermodynamics relationships between energy (ε), Helmholtz free energy (f) and entropy (s), which implies [28]:

$$\Delta\varepsilon_{bs} = \Delta f_{bs} - T\frac{d(\Delta f_{bs})}{dT}, \tag{5.4}$$

where Δf_{bs} is the free energy change for moving one molecule from the bulk of liquid to the surface. Another straightforward way to obtain Eq. (5.4) is using the Gibbs-Helmholtz equation:

$$\frac{d(\frac{f}{T})}{dT} = -\frac{\varepsilon}{T^2}, \tag{5.5}$$

which implies

$$\frac{1}{T}\frac{df}{dT} - \frac{f}{T^2} = -\frac{\varepsilon}{T^2}. \tag{5.6}$$

This then clearly yields Eq. (5.4) by subtracting the resulting equations for the bulk and surface states.

Exploiting the thermodynamics relationship among Δf_{bs}, surface tension (γ) and the molecular surface area (a_s) which is defined as [28]:

$$\Delta f_{bs} = \frac{a_s}{2}\,\gamma. \tag{5.7}$$

Eq. (5.4) can be rewritten as:

$$\Delta\varepsilon_{bs} = \frac{a_s}{2}(\gamma - T\frac{d\gamma}{dT}). \tag{5.8}$$

Halving the molecular surfaces a_s in Eq. (5.7) is considered here because in fact only one-half of the molecular surfaces contribute to forming the gas-liquid interface and the other half remains in the bulk of the liquid [28].

By substituting Eq. (5.8) in Eq. (5.3) and using $\frac{d(\frac{\gamma}{T})}{dT} = \frac{T\frac{d\gamma}{dT}-\gamma}{T^2}$ and multiplying both sides by Avogadro's number, the correlation between the surface tension and molar vaporization enthalpy (ΔH_{vap}) is obtained as :

$$\Delta H_{vap} = \frac{a_s}{2} N_A \left(2\,\gamma - T\,\frac{d\gamma}{dT}\right) - \frac{R}{2}\,T\ln(T) + \beta T\,, \tag{5.9}$$

in which β is a constant. Knowing that at the critical temperature both vaporization enthalpy and surface tension approach zero, and due to continuity of the surface tension, the $\frac{d\gamma}{dT}$ term also approaches zero, the constant β is found as:

$$\beta = \frac{R}{2}\ln(T_c)\,, \tag{5.10}$$

which by substitution into Eq. (5.9) finally results in:

$$\Delta H_{vap} = \frac{a_s}{2}\left(2\,\gamma - T\,\frac{d\gamma}{dT}\right) - \frac{R}{2}\,T\ln\left(\frac{T}{T_c}\right). \tag{5.11}$$

As discussed earlier, the main advantage of the theoretically derived relationship among vaporization enthalpy, surface tension, and molecular surfaces described by Eq. (5.11) is the possibility of experimental determination of molecular surfaces through this relationship as well as evaluation of the currently defined molecular surfaces. For the latter goal, we have compared the accuracy of evaluated vaporization enthalpies obtained via Eq. (5.11) using 252 differently computed molecular surfaces based on various parameterizations of vdW and SAS methods.

5.3 Experimental

5.3.1 Dataset

The theoretically derived methods were benchmarked using thermophysical data of the DIPPR801 database [29]. Screening the initial dataset and selecting only the compounds with a maximum uncertainty of 5% for vaporization enthalpy and surface tension resulted in 215 compounds from diverse chemical families, provided as supplementary materials.

For each compound, the experimentally determined data of vaporization enthalpies for 25 temperatures linearly distributed between the melting point and the critical temperature were

taken using the provided relationships in the DIPPR database. Due to the scarcity of accurate surface tension data for all of the required data points at which vaporization enthalpy data were available, we employed the Guggenheim–Katayama relationship stated as [30]:

$$\gamma = \gamma^{\circ}\left(1 - \frac{T}{T_c}\right)^{11/9}, \tag{5.12}$$

as a rigorous model for evaluating temperature dependence of surface tension and its temperature derivatives required by Eq. (5.11). One main advantage of employing the Guggenheim–Katayama relationship is that it perfectly satisfies the boundary conditions, i.e. yielding exactly zero for surface tension and its higher-order derivatives with respect to temperature at the critical point, which is commonly violated by other relationships like those proposed in the DIPPR dataset. This is indeed a key requirement, as it was one of the premises of obtaining the constant β in Eq. (5.11). Additionally, compared to various surface tension predictive models such as those used by the DIPPR database and the Eötvös relationship [18], we found that the Guggenheim–Katayama relationship provides the most accurate evaluation of not only temperature dependence of surface tension but also its derivate for the whole temperature range and hence, the most accurate prediction of vaporization enthalpy via Eq. (5.11).

To obtain surface tension data at the required temperatures via the Guggenheim–Katayama relationship, for each compound the pre-factor γ° was calculated by optimization using the available experimental data points of surface tension. The calculated values of γ° for each compound are reported in the supplementary materials.

The accuracy of the predicted vaporization enthalpy is reported as Average Absolute Deviation (AAD), defined as:

$$\mathrm{AAD} = \frac{1}{N}\sum\left(\left|y_i^{exp} - y_i^{pred}\right|\right). \tag{5.13}$$

Considering that at the critical point, vaporization enthalpy approaches zero and slight deviations in predicted enthalpies results in very large relative errors, AAD provides a more reliable evaluation of the model performances, compared to relative errors.

5.3.2 Computational details

To calculate well-established molecular surfaces, the geometry of each molecule was first optimized at the B3LYP/6-311+G (2d,p) level of theory. Using the optimized structures, the

vdW molecular surfaces and Solvent Accessible Surface Areas (SASA) for each molecule were calculated using the GEPOL algorithm implemented in the Gaussian 16 software [31].

For both vdW and SASA, we have calculated molecular surfaces using UA0, Pauling, Bondi, and UFF parameterizations of atomic radii, as available in the Gaussian 16 software [31]. For each one of the mentioned parameterizations, we have also studied the atomic radii scaled by factors 0.85, 0.9, 0.95, 1.0, 1.05, 1.1, 1.15, 1.2, and 1.25.

For SASA, we computed solvent radii required for calculation of solvent accessible surfaces using radii of perfect spheres with volumes equal to the vdW volume and also molar volumes of the solvents at their melting points. For each one of the employed solvent radii, in addition to the original values, radii scaled by factors of 0.7 and 0.85 were also studied. The reason of such scaling is that we noticed for most of the studied molecules, the radii computed via vdW or melting point volumes deviate from the respective solvent radii used in Gaussian 16 in the calculation of continuum solvation models by a factor varying between these two values.

With all this diversity in details, in total 252 differently computed molecular surfaces, in addition to molecular surfaces approximated based on the perfect sphere assumption and molar volumes at melting point and normal boiling point, were studied for each molecule, and their performance in reproducing experimentally determined vaporization enthalpy via Eq. (5.11) was tested.

5.4 Results and discussion

5.4.1 Verifying the validity of the theoretically derived relationship

By studying various relationships proposed in the past century to analytically relate solution thermodynamics, surface tension, and molecular surfaces, we noticed an obvious inconsistency not only between previously proposed models themselves but also with the theoretically derived relationship proposed in the present study as well. Accordingly, while a large number of studies [19-22,25,32,33] support:

$$\Delta G_{solvation} = \mathcal{A}\,\gamma + \mathcal{B}, \tag{5.14}$$

where $\Delta G_{solvation}$ is the free energy of solvation, $\mathcal{A}$ is the solvent excluded surface of molecules and $\mathcal{B}$ is a constant [34], many other works employ the very similar relationship to relate vaporization enthalpy, surface tension, and molecular surfaces. A well-known example of the latter category is Kabo's method which proposes:

$$\Delta H_{vap} = \mathcal{A}\left(V^{\frac{2}{3}} \gamma \right) + \mathcal{B}, \tag{5.15}$$

and has been extensively applied especially in studying the phase equilibrium in ionic liquids [35]. Noteworthy, the right-hand side of Kabo's relationship closely matches the right-hand side of Eq. (5.14) with the only difference that in Kabo's relationship, the pre-factor of the surface tension is in fact molecular surfaces evaluated based on the perfect sphere assumption discussed earlier, while in Eq. (5.14) this factor is solvent excluded surface of molecules [34].

Alongside the mentioned paradoxical deviations between the two models, both of them also show obvious inconsistencies compared to our theoretically derived relationship proposed in Eq. (5.11) which implies the necessity of a careful and rigorous verification of our model.

To that end, we first studied the overall accuracy of vaporization enthalpies predicted via the newly developed relationship. Accordingly, for each compound, we optimized the value of the a_s parameter required by Eq. (5.11) which yielded the lowest error in predicting vaporization enthalpies over the whole temperature range. Via the optimized a_s parameters, which are in fact our proposed thermodynamically effective molecular surfaces, an AAD of 0.188 kcal/mol was obtained for the predicted vaporization enthalpies of the whole dataset. This resulting AAD is within both the chemical accuracy and the reported accuracy of the experimentally determined data.

Interestingly, we noticed that the molecular surfaces approximated via molar volumes at the melting point based on the perfect sphere assumption divided by the factor of two to include only the part of the molecular surfaces which contribute to the interface as discussed earlier, closely match the thermodynamically effective molecular surfaces with an average ratio of 1.0372, standard deviation of 0.117 and correlation coefficient of 0.95. These results clearly verify the robustness of the theoretically derived relationship and its underlying presumptions like defining the a_s parameters as molecular surfaces at the interface and considering the factor of ½ for it.

Noteworthy, using the molecular surfaces evaluated from molar volumes at the melting point and without any adjustable parameters, Eq. (5.11) yielded an AAD of 0.561 kcal/mol for the whole dataset and the whole temperature range. On the other hand, using molecular surfaces obtained via molar volumes at normal boiling points resulted in AAD of 1.248 kcal/mol. These results clearly show that the perfect sphere assumption is not generally a rigorous approach for evaluating molecular surfaces. In addition to lower accuracy in describing temperature dependence of vaporization enthalpy compared to the thermodynamically

effective molecular surfaces, it totally overlooks temperature dependence of molar volume and its impact on the evaluated molecular surfaces.

To investigate the significance of the $T\frac{d\gamma}{dT}$ and $-\frac{R}{2}\,T\ln\left(\frac{T}{T_c}\right)$ terms, which are the most obvious differences between our proposed relationship and the conventionally accepted models, we re-optimized a_s parameters for the two following relationships:

$$\Delta H_{vap} = a_s\gamma - \frac{R}{2}\,T\ln\left(\frac{T}{T_c}\right), \tag{5.16}$$

and

$$\Delta H_{vap} = \frac{a_s}{2}\left(2\,\gamma - T\,\frac{d\gamma}{dT}\right), \tag{5.17}$$

which are two variants of Eq. (5.11) obtained by removing the term being studied. Using the re-optimized a_s parameters, the two abovementioned variants yielded AADs of 2.071 and 0.197 kcal/mol, respectively. These results clearly show that both terms $\frac{d\gamma}{dT}$ and $-\frac{R}{2}\,T\ln\left(\frac{T}{T_c}\right)$ play a significant role in describing the relationship between solution thermodynamics and molecular surfaces. Among them, the $T\,\frac{d\gamma}{dT}$ term has the more significant impact and its overlooking reduces the accuracy of predicted vaporization enthalpies by one order of magnitude.

5.4.2 Evaluation of molecular surfaces estimated via computer algorithms

After verifying the validity and robustness of the theoretically derived relationship in the previous section, this section focuses on studying the predictability of vaporization enthalpies obtained via various parameterizations of the vdW and solvent accessible surfaces. To that end, we have studied a total number of 252 differently computed molecular surfaces discussed in section 5.3.2.

According to the results, while via the solvent accessible surfaces we could not achieve any AAD better than 7.889 kcal/mol, for the vdW surfaces the best results with AAD of 0.568 kcal/mol was obtained for UA0 atomic radii scaled by 0.9. For the vdW surfaces, the results obtained for original parameterizations of atomic radii without any scaling yielded AADs of 0.613, 1.078, 1.256, and 0.867 kcal/mol for Bondi, UFF, UA0, and Pauling parameterizations, respectively.

These results show that which method is selected for approximating molecular surfaces clearly has a remarkable impact on the thermodynamic quantities evaluated with those surfaces. For a better illustration, the AADs of predicted vaporization enthalpies obtained via Eq. (5.11) for different scalings and parameterizations of atomic radii are compared in figure 5.2. As can be seen in figure 5.2, even slight differences in approximated molecular surfaces can have a significant impact on the evaluated vaporization enthalpies in solution.

Noteworthy, the vdW surfaces with UA0 atomic radii scaled by 0.9, which yielded better results compared to other parameterizations for vaporization enthalpy prediction, after scaling by the ½ factor as discussed earlier, show an excellent agreement with the calculated thermodynamically effective surfaces, with a correlation coefficient of 0.96, the average ratio of 1.019 and standard deviation of 0.172.

This immediately suggests employing thermodynamically effective surfaces for a more rigorous parameterization of atomic radii in the currently defined molecular surfaces.

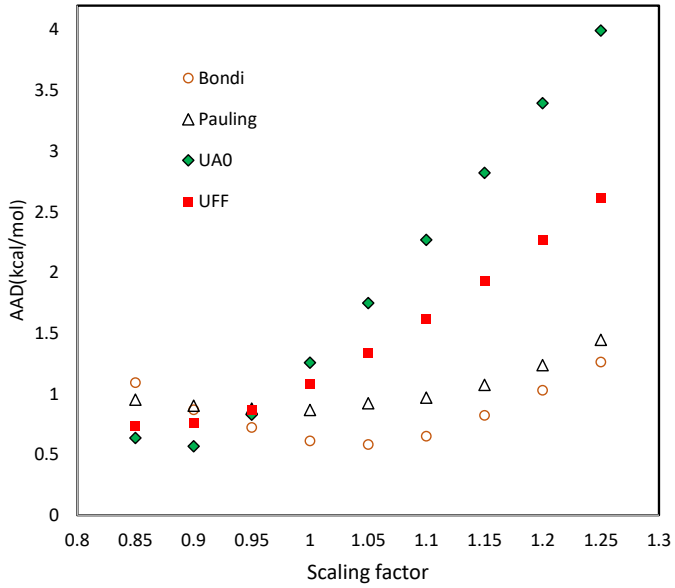


Fig. 5.2 Comparison of AAD of predicted vaporization enthalpies for various parametrizations of atomic radii scaled by alpha constant in vdW model.

In figure 5.4, the temperature dependence of vaporization enthalpy of some of the most widely used solvents predicted via Eq. (5.11) using various molecular surfaces are compared. These results clearly show the significant importance of evaluated molecular surfaces on the accuracy of obtained results on the one hand and the robustness and reliability of the newly

derived relationship, and the thermodynamically effective molecular surfaces on the other hand.

As discussed in section 5.4.1, there are paradoxical inconsistencies between our proposed relationship for describing the dependency of solution thermodynamics to molecular surfaces with those proposed in the literature. To compare our proposed relationship with conventionally accepted models, we evaluated the accuracy of vaporization enthalpies predicted via Kabo's relationship, for which the two constants were determined by optimization using the whole vaporization enthalpy data for each compound. According to the results, while the newly derived relationship with optimum molecular surfaces yielded AAD of 0.188 kcal/mol, Kabo's method with two empirical parameters optimized for each compound could not yield an AAD better than 0.402 kcal/mol which is comparable to the results obtained via molecular surfaces at the melting point without needing any adjustable parameters, as reported earlier. The same comparison of our proposed relationship and Eq. (5.14) as the other conventionally accepted model for relating solution thermodynamics to molecular surfaces is provided in section 5.4.3.

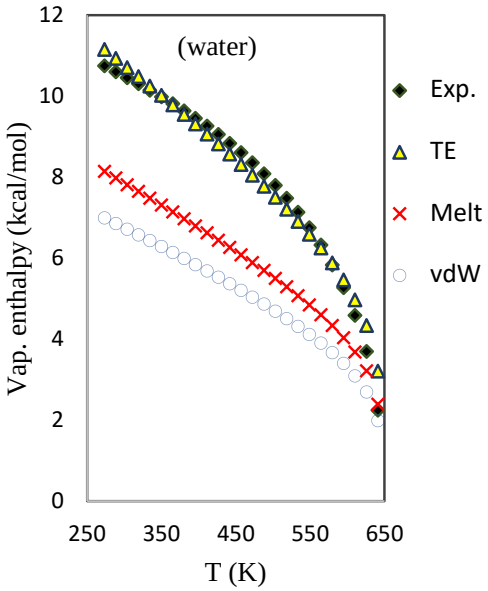
5.4.3 Estimation of other thermodynamics quantities via thermodynamically effective surfaces

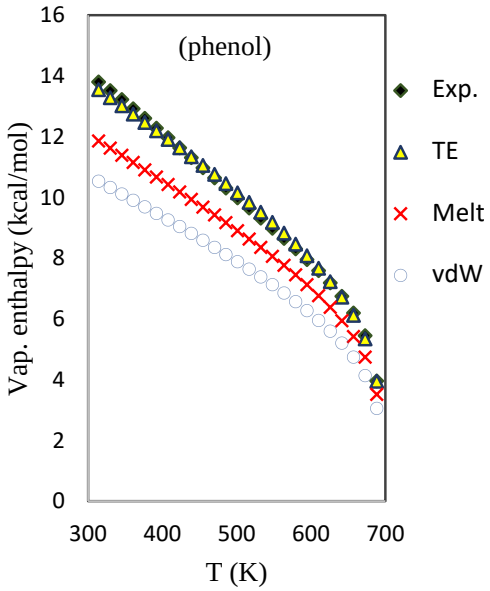
As it was discussed earlier, one of the main advantages of characterizing dependency of solution thermodynamics on molecular surfaces through temperature dependence of vaporization enthalpy is its straightforward transferability to other thermodynamics quantities.

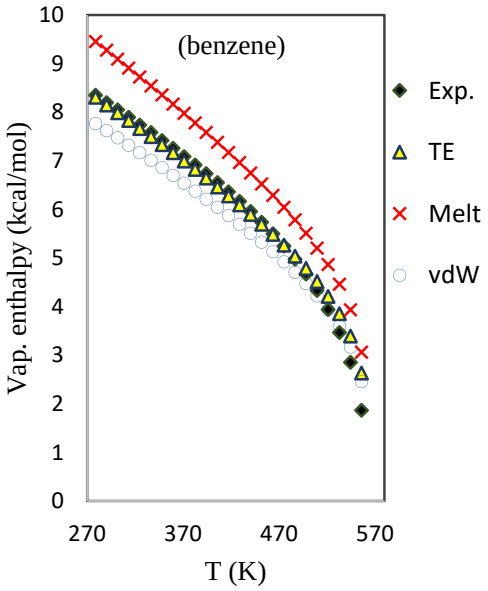
As one of the most important thermodynamics quantities, evaluation of the solvation free energy, which is the primary goal in continuum solvation models [10], can be achieved via the Gibbs-Helmholtz relationship as follows:

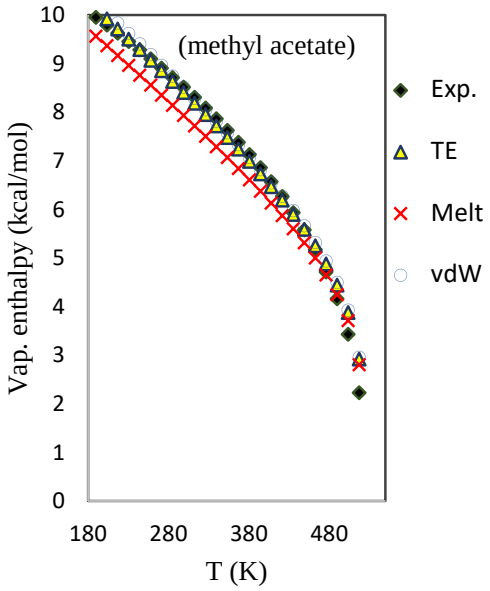
$$\Delta G_{solvation} = -T \int \frac{\Delta H_{vap}}{T^2}\, dT\,. \tag{5.18}$$

Via the analytical relationship describing temperature dependence of vaporization enthalpy, one can solve Eq. (5.18) analytically or numerically. The constant of integration then needs to be determined using a reference data point. To that end, the most convenient choice is to use the free energy of solvation at the normal boiling point, where the saturation vapor pressure of the liquid becomes equal to the atmospheric pressure. It then allows to determine the solvation Gibbs free energy at the normal boiling point through atmospheric pressure (P_{atm}), liquid molar volume at the normal boiling point ($V_{m,nbp}$), and normal boiling point temperature (T_{nbp}) via [36]:


(water)
Vap. enthalpy (kcal/mol)
T (K)
Exp.
TE
Melt
vdW


(phenol)
Vap. enthalpy (kcal/mol)
T (K)
Exp.
TE
Melt
vdW


(benzene)
Vap. enthalpy (kcal/mol)
T (K)
Exp.
TE
Melt
vdW


(methyl acetate)
Vap. enthalpy (kcal/mol)
T (K)
Exp.
TE
Melt
vdW

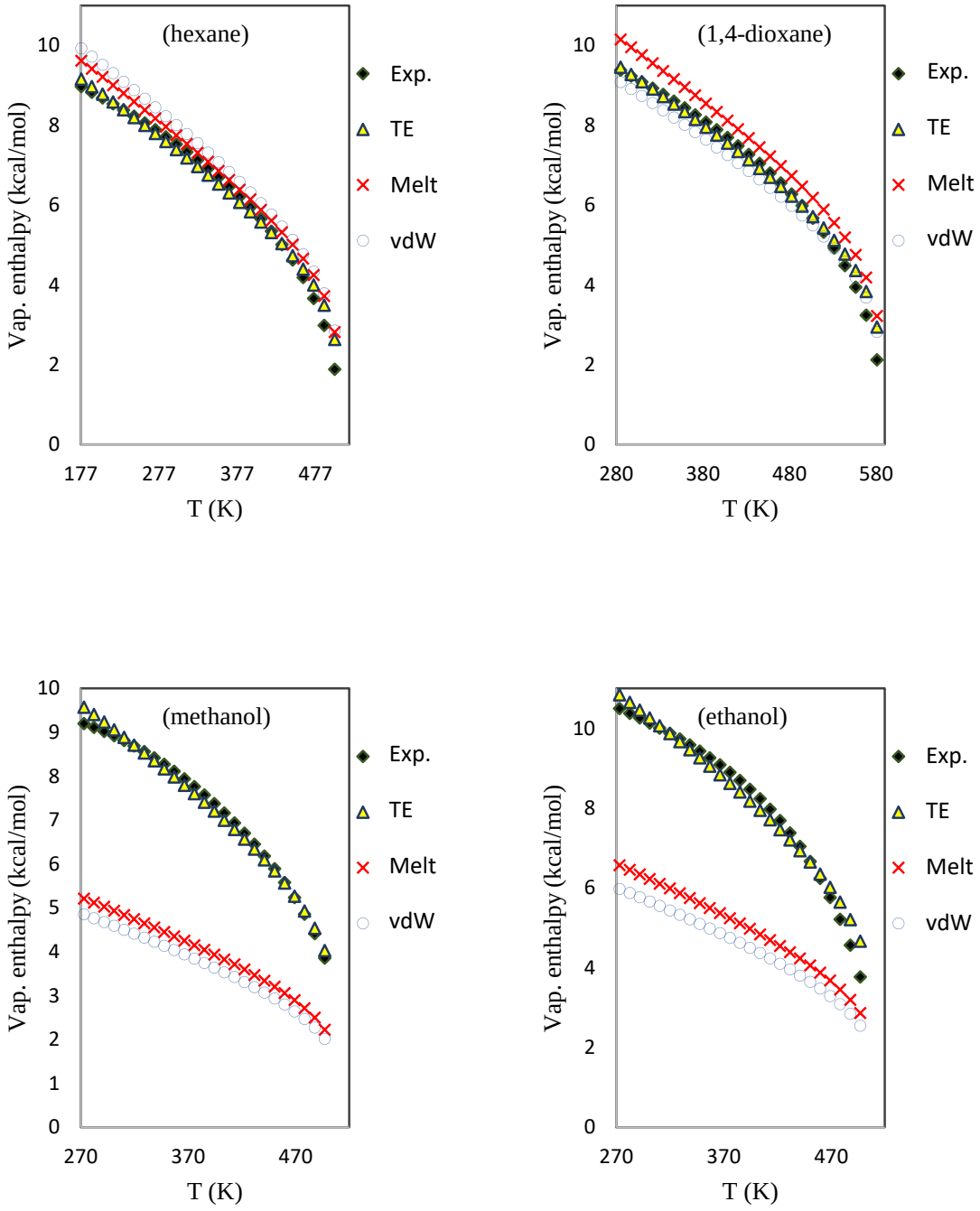


Fig. 5.4 Comparison of the impact of various molecular surfaces on the accuracy of predicted vaporization enthalpies. The results show a remarkably enhanced description of vaporization enthalpy at various temperatures using our newly proposed thermodynamically effective (TE) surfaces compared to the surfaces evaluated from melting point volumes (melt) and vdW surfaces obtained via UA0 parameterization and scaled by 0.9.

$$P_{atm} = \frac{R\,T_{nbp}}{V_{m,nbp}} \exp\left(\frac{\Delta G_{solvation,nbp}}{RT_{nbp}}\right). \tag{5.19}$$

The abovementioned dependency between vaporization enthalpy and solvation free energy also allows a quantitative comparison of our theoretically derived relationship with the other conventionally accepted model, which in contrast to our model suggests linear dependency between solvation thermodynamics and the product of molecular surfaces and surface tension via Eq. (5.14). To that end, we studied the predictability of the solvation free energies obtained by numerically integrating the Gibbs-Helmholtz equation for the experimentally determined vaporization enthalpies as a function of temperature and used the obtained free energies to find the parameter $\mathcal{B}$ in Eq. (5.14) by optimization. For the solvation free energies predicted this way, the results showed an AAD of 1.166 kcal/mol which is again much higher than the AAD of computed vaporization enthalpies. Noteworthy, employing the newly proposed relationship for numerically integrating the Gibbs-Helmholtz equation could reproduce the reference solvation free energies with AAD of only 0.057 kcal/mol. The much higher accuracy obtained via thermodynamically effective surfaces for predicted solvation free energies is because the solvation free energies are proportional to the area under the curve of vaporization enthalpy versus temperature and therefore are less affected by outliers. These results also suggest that considering linear proportionality between vaporization enthalpy and surface tension, which has already been empirically suggested in numerous studies like Kabo's relationship and is supported in the present study, is more valid than the other conventionally employed relationships which consider the same proportionality between solvation free energy and surface tension.

We also studied the predictability of experimentally determined standard state solvation free energies reported in the Minnesota Solvation database [37] via thermodynamically effective surfaces for the solutions common between the DIPPR and Minnesota Solvation databases. The results, which are reported in table 5.1, show an excellent agreement between predicted solvation free energies obtained via the thermodynamically effective surfaces and the experimentally determined values reported in the Minnesota Solvation database. The obtained accuracy of these results shows an AAD of only 0.1215 kcal/mol which is by almost a factor of 2 more accurate than best results obtained via advanced continuum solvation models [10].

In addition to the solvation free energy, the saturation vapor pressure (P^{sat}) as another extensively required thermodynamic quantity in many industrial and scientific applications can

Table 5.1 Comparison of standard state solvation free energies theoretically predicted using thermodynamically effective surfaces and experimentally determined data.

Compound	Exp. (kcal/mol)	Predicted (kcal/mol)
Acetonitrile	-4.85	-4.9191
2-methylpyridine	-5.71	-5.8055
Acetophenone	-7.59	-7.7809
Aniline	-7.61	-7.5768
Anisole	-6.33	-6.4320
Benzene	-4.55	-4.5595
Benzonitrile	-7.28	-7.4975
Bromoethane	-3.67	-3.7014
Bromoform	-6.21	-6.3047
1-butanol	-6.03	-6.1576
Chlorobenzene	-5.66	-5.7420
Chloroform	-4.13	-4.1932
Cyclohexane	-4.43	-4.4176
Cyclohexanone	-6.25	-6.3928
1-decanol	-9.58	-9.3872
N,N-dimethylacetamide	-6.77	-6.8710
2,6-dimethylpyridine	-6.04	-6.0984
Ethanol	-5.04	-5.0844
Ethylbenzene	-5.67	-5.7657
Fluorobenzene	-4.60	-4.6416
1-heptanol	-7.84	-7.7992
1-hexanol	-7.05	-7.2207
2-Methyl-1-Propanol	-5.79	-5.8164
2-Propanol	-4.82	-5.0727
Isopropylbenzene	-6.04	-6.1356
m-cresol	-8.40	-8.3327
Mesitylene	-6.40	-6.5034
dichloromethane	-3.80	-3.8348
Nitrobenzene	-7.94	-8.0294
Nitroethane	-5.53	-5.6490
Nitromethane	-5.38	-5.4908
1-nonanol	-9.05	-8.9186
1-pentanol	-7.92	-6.6836
1-propanol	-5.29	-5.5662
Pyridine	-5.47	-5.5464
2-Butanol	-5.48	-5.5590
tert-Butylbenzene	-6.43	-6.4638
Tetrahydrofuran	-4.25	-4.2791
Toluene	-5.12	-5.1760
Triethylamine	-4.44	-4.4499
1,2,4-Trimethyl Benzene	-6.47	-6.6175
Water	-6.31	-6.4197
AAD	0.1215 kcal/mol	

also be accurately computed via the thermodynamically effective surfaces via the Clausius–Clapeyron relation, which for ideal gases can be written as:

$$\ln\left(\frac{P^{sat}}{P_{atm}}\right) = \int_{T_{nbp}}^{T} \frac{\Delta H_{vap}}{R\,T^2}\,dT\,. \tag{5.20}$$

The saturation vapor pressure of non-ideal gases can also be determined the same way via evaluating the temperature dependence of gas-phase molar volumes using an appropriate equation of states. Nevertheless, in the present study and only for a proof of concept, we study the predictability of saturation vapor pressures of ideal gases (pressures up to 2.5 atm) via Eq. (5.20). A comparison of predicted and experimentally determined saturation vapor pressures for a number of most widely used solvents is depicted in figure 5.5. The excellent agreement between the theoretically evaluated and experimental data depicted in figure 5.5 implies the robustness of our proposed thermodynamically effective surfaces.

5.5 Conclusion

In summary, in the context of the present study, we could theoretically derive a relationship that describes dependency between vaporization enthalpy, molecular surfaces, and surface tension. We could demonstrate that the proposed dependency between solution thermodynamics and molecular surfaces is remarkably more reliable and rigorous compared to the other models empirically proposed for the same purpose within the last century.
Through our newly derived theoretical approach, we proposed thermodynamically effective surfaces. We demonstrated that the thermodynamically effective surfaces are only slightly different than empirically proposed vdW surfaces. However, this slight deviation yields a substantial improvement in the predictability of multiple thermodynamic quantities. As a result, we propose the thermodynamically effective surfaces as a reliable alternative for the currently defined molecular surfaces, especially for studying the condensed phase thermodynamics.

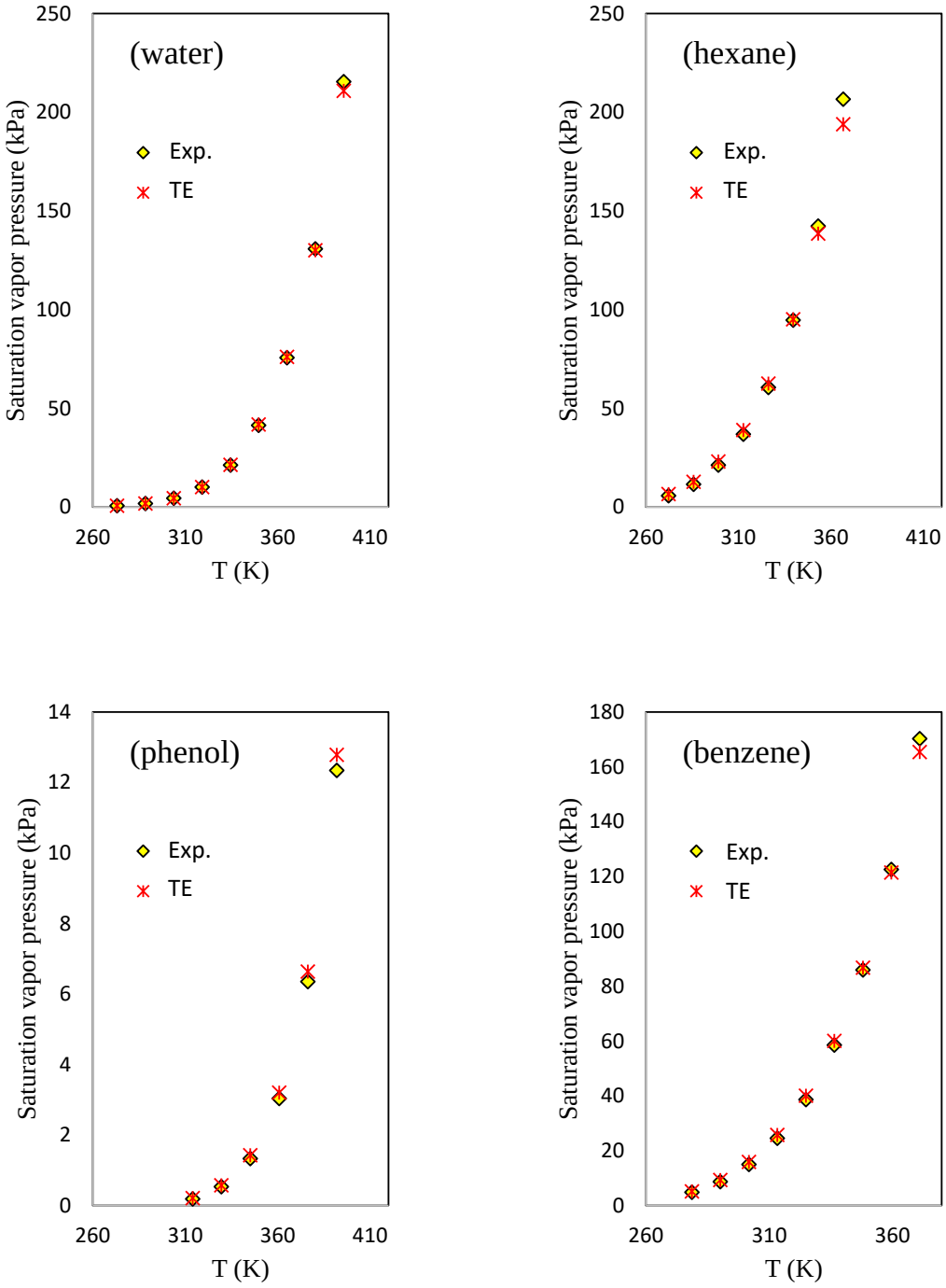


Fig. 5.5 Comparison of theoretically predicted saturation vapor pressures obtained via thermodynamically effective surfaces (denoted by TE) with experimentally determined data.

References:

1. Zimmerman, M. I.; Porter, J. R.; Ward, M. D.; Singh, S.; Vithani, N.; Meller, A.; Mallimadugula, U. L.; Kuhn, C. E.; Borowsky, J. H.; Wiewiora, R. P., SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nature Chemistry 2021, 1-9.
2. Portelli, S.; Phelan, J. E.; Ascher, D. B.; Clark, T. G.; Furnham, N., Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Scientific reports 2018, 8 (1), 1-12.
3. Christoffel, F.; Igareta, N. V.; Pellizzoni, M. M.; Tiessler-Sala, L.; Lozhkin, B.; Spiess, D. C.; Lledós, A.; Maréchal, J.-D.; Peterson, R. L.; Ward, T. R., Design and evolution of chimeric streptavidin for protein-enabled dual gold catalysis. Nature Catalysis 2021.
4. Lee, J.; Chang, I.; Yu, W., Atomic insights into the effects of pathological mutants through the disruption of hydrophobic core in the prion protein. Scientific reports 2019, 9 (1), 1-14.
5. Mishra, A.; Ranganathan, S.; Jayaram, B.; Sattar, A., Role of solvent accessibility for aggregation-prone patches in protein folding. Scientific reports 2018, 8 (1), 1-13.
6. Gristick, H. B.; von Boehmer, L.; West Jr, A. P.; Schamber, M.; Gazumyan, A.; Golijanin, J.; Seaman, M. S.; Fätkenheuer, G.; Klein, F.; Nussenzweig, M. C., Natively glycosylated HIV-1 Env structure reveals new mode for antibody recognition of the CD4-binding site. Nature structural & molecular biology 2016, 23 (10), 906-915.
7. Waskiewicz, E.; Vasiliou, M.; Corcoles-Saez, I.; Cha, R. S., Cancer genome datamining and functional genetic analysis implicate mechanisms of ATM/ATR dysfunction underpinning carcinogenesis. Communications biology 2021, 4 (1), 1-11.
8. Van der Verren, S. E.; Van Gerven, N.; Jonckheere, W.; Hambley, R.; Singh, P.; Kilgour, J.; Jordan, M.; Wallace, E. J.; Jayasinghe, L.; Remaut, H., A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. Nature biotechnology 2020, 38 (12), 1415-1420.
9. Gainza, P.; Sverrisson, F.; Monti, F.; Rodola, E.; Boscaini, D.; Bronstein, M.; Correia, B., Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods 2020, 17 (2), 184-192.
10. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. Nature Communications 2021,12(1),1-7.
11. Kim, D. H.; Ringe, S.; Kim, H.; Kim, S.; Kim, B.; Bae, G.; Oh, H.-S.; Jaouen, F.; Kim, W.; Kim, H., Selective electrochemical reduction of nitric oxide to hydroxylamine by atomically dispersed iron catalyst. Nature communications 2021, 12 (1), 1-11.
12. Guo, L.; Srimontree, W.; Zhu, C.; Maity, B.; Liu, X.; Cavallo, L.; Rueping, M., Nickel-catalyzed Suzuki–Miyaura cross-couplings of aldehydes. Nature communications 2019, 10 (1), 1-6.
13. Liu, Y.-M.; Hou, H.; Zhou, Y.-Z.; Zhao, X.-J.; Tang, C.; Tan, Y.-Z.; Müllen, K., Nanographenes as electron-deficient cores of donor-acceptor systems. Nature communications 2018, 9 (1), 1-7.

14. Rousseau, R.; Glezakou, V.-A.; Selloni, A., Theoretical insights into the surface physics and chemistry of redox-active oxides. Nature Reviews Materials 2020, 5 (6), 460-475.
15. Biasin, E.; Fox, Z. W.; Andersen, A.; Ledbetter, K.; Kjær, K. S.; Alonso-Mori, R.; Carlstad, J. M.; Chollet, M.; Gaynor, J. D.; Glownia, J. M., Direct observation of coherent femtosecond solvent reorganization coupled to intramolecular electron transfer. Nature Chemistry 2021, 13 (4), 343-349.
16. Han, Y.-X.; Jiang, Y.-L.; Li, Y.; Yu, H.-X.; Tong, B.-Q.; Niu, Z.; Zhou, S.-J.; Liu, S.; Lan, Y.; Chen, J.-H., Biomimetically inspired asymmetric total synthesis of (+)-19-dehydroxyl arisandilactone A. Nature communications 2017, 8 (1), 1-13.
17. Kim, T. Y.; Jeon, S.; Jang, Y.; Gotina, L.; Won, J.; Ju, Y. H.; Kim, S.; Jang, M. W.; Won, W.; Park, M. G., Platycodin D, a natural component of Platycodon grandiflorum, prevents both lysosome-and TMPRSS2-driven SARS-CoV-2 infection by hindering membrane fusion. Experimental & molecular medicine 2021, 53 (5), 956-972.
18. Eötvös, R., Ueber den Zusammenhang der Oberflächenspannung der Flüssigkeiten mit ihrem Molecularvolumen. Annalen der Physik 1886, 263 (3), 448-459.
19. Sisskind, B.; Kasarnowsky, I., Untersuchungen über die Löslichkeit der Gase. 2. Mitteilung. Löslichkeit des Argons. Zeitschrift für anorganische und allgemeine Chemie 1933, 214 (4), 385-395.
20. Uhlig, H., The solubilities of gases and surface tension. Journal of Physical Chemistry 1937, 41 (9), 1215-1226.
21. Eley, D., On the solubility of gases. Part I.—The inert gases in water. Transactions of the Faraday Society 1939, 35, 1281-1293.
22. Clever, H. L.; Battino, R.; Saylor, J.; Gross, P., The solubility of helium, neon, argon and krypton in some hydrocarbon solvents. The Journal of Physical Chemistry 1957, 61 (8), 1078-1082.
23. Bondi, A. v., van der Waals volumes and radii. The Journal of physical chemistry 1964, 68 (3), 441-451.
24. Gallicchio, E.; Kubo, M.; Levy, R. M., Enthalpy-entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation. The Journal of Physical Chemistry B 2000, 104 (26), 6271-6285.
25. Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K., On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute-solvent interaction energy. Journal of the American Chemical Society 2003, 125 (31), 9523-9530.
26. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., Gaussian 16. Revision A 2016, 3.
27. Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M., UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. Journal of the American chemical society 1992, 114 (25), 10024-10035.
28. Alibakhshi, A., Enthalpy of vaporization, its temperature dependence and correlation with surface tension: a theoretical approach. Fluid Phase Equilibria 2017, 432, 62-69.
29. Wilding, W. V.; Rowley, R. L.; Oscarson, J. L., DIPPR® Project 801 evaluated process design data. Fluid phase equilibria 1998, 150, 413-420.
30. Adam, N., he physics and chemistry of surfaces (3d ed.): Oxford University Press. London: 1941.
31. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., Gaussian 16. Gaussian, Inc. Wallingford, CT: 2016.
32. Reynolds, J. A.; Gilbert, D. B.; Tanford, C., Empirical correlation between hydrophobic free energy and aqueous cavity surface area. Proceedings of the National Academy of Sciences 1974, 71 (8),

2925-2927.
33. Hermann, R. B., Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions. Proceedings of the National Academy of Sciences 1977, 74 (10), 4144-4145.
34. Zacharias, M., Continuum solvent modeling of nonpolar solvation: Improvement by separating surface area dependent cavity and dispersion contributions. The Journal of Physical Chemistry A 2003, 107 (16), 3000-3004.
35. Zaitsau, D. H.; Kabo, G. J.; Strechan, A. A.; Paulechka, Y. U.; Tschersich, A.; Verevkin, S. P.; Heintz, A., Experimental vapor pressures of 1-alkyl-3-methylimidazolium bis (trifluoromethylsulfonyl) imides and a correlation scheme for estimation of vaporization enthalpies of ionic liquids. The Journal of Physical Chemistry A 2006, 110 (22), 7303-7306.
36. Akkermans, R. L., Solvation Free Energy of Regular and Azeotropic Molecular Mixtures. The Journal of Physical Chemistry B 2017, 121 (7), 1675-1683.
37. Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G., Minnesota solvation database. Minnesota Solvation Database version 2012, 20.

Chapter 6

Publication: Strategies to develop rigorous ANN models

6.1 Scope of the Project

6.1.1 Project overview and motivation:

Artificial neural networks are one of the most powerful and most widely applied machine-learning approaches in theoretical chemistry. The availability of numerous user-friendly software tools has made employing artificial neural networks a convenient task, even without requiring profound knowledge on its theoretical background. Nevertheless, a downside of this ease of application without studying the theoretical background is overlooking some important intricacies which can soon result in the unreliability of the models developed this way. In this work, I introduced the most important intricacies which can be considered as the minimal knowledge every scientist who is interested in developing neural network models should be aware of and carefully consider. These intricacies include the appropriate selection of the training algorithms, transfer functions, and more importantly the number of neurons. I provided validation strategies that can significantly enhance the reliability of the developed neural network models. I benchmarked the provided guidelines by studying the predictability of flash-point for a large data set of hydrocarbons based on the group contribution and neural networks.

6.1.2 Novelty aspects:

- The provided rule of thumb for evaluating the upper limit of hidden later neurons

- The validation strategy which is beyond the conventional cross-validation and employs rigorous statistical inferring
- Prediction of the flash-points for the studied compounds via the developed machine learning models yielded the most accurate reported results until then

6.1.3 Connection to other chapters:

The provided guidelines and strategies are the cornerstones of the presented machine learning models in chapters 7 to 9.

6.2 Publication Data and Reprint

Reference:	Alibakshi, A., Strategies to develop robust neural network models: Prediction of flash point as a case study. *Analytica chimica acta* **2018,** *1026*, 69-76.
Submitted:	7 August 2017
Accepted:	3 May 2018
Contribution:	Carrying out all the computations, method development, and writing the manuscript
Copyright:	Reproduced with permission from the journal of "Analytica Chimica Acta". Copyright 2018 Elsevier B.V.

Analytica Chimica Acta 1026 (2018) 69–76

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Strategies to develop robust neural network models: Prediction of flash point as a case study

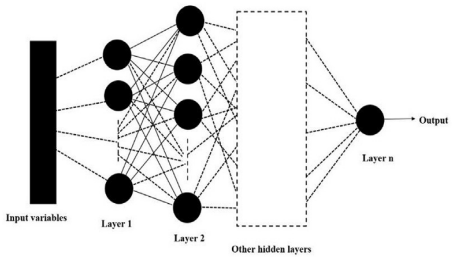
Amin Alibakshi

Department of Marine Geosystems, Geomar Helmholtz Center for Ocean Research Kiel, Wischhofstrasse 1-3, 24148, Kiel, Germany

HIGHLIGHTS

- The present study introduces the important practical aspects of developing a reliable artificial neural network (ANN) model including appropriate assignment of the number of neurons, number of hidden layers, transfer functions, training algorithms, dataset division and initialization of the network.
- A rule of thumb is suggested for evaluating the appropriate number of neurons and layers (the ratio of the training samples to the ANN constants should be equal or greater than 10).
- A strategy for evaluating of the authentic performance of an ANN model as well the appropriateness of training is proposed.
- Using the introduced considerations, a model is developed for predicting the flash point of pure compounds which produces the lowest error compared to other models.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 7 August 2017
Received in revised form
2 May 2018
Accepted 3 May 2018
Available online 9 May 2018

Keywords:
Artificial neural networks
Supervised learning
Overfitting
Appropriate training
Group contribution method
QSPR

ABSTRACT

Artificial neural network (ANN) is one of the most widely used methods to develop accurate predictive models based on artificial intelligence and machine learning. In the present study, the important practical aspects of developing a reliable ANN model e.g. appropriate assignment of the number of neurons, number of hidden layers, transfer function, training algorithm, dataset division and initialization of the network are discussed. As a case study, predictability of the flash point for a dataset of 740 organic compounds using ANNs was investigated via a total number of 484220ANNs to allow covering a wide range of parameters affecting the performance of an ANN. Among all studied parameters, the number of neurons or layers was found to be the most important parameters to develop a reliable ANN with low overfitting risk. To evaluate appropriate number of neurons and layers, a value of equal or greater than 10 for the ratio of the training samples to the ANN constants was suggested as a rule of thumb. More ever, a strategy for evaluation of the authentic performance of ANNs and deciding about the reliability of an ANN model was proposed which is applicable to other models developed by supervised learning.

E-mail addresses: aalibakhshi@geomar.de, Amin.alibakhshi@hotmail.com.

https://doi.org/10.1016/j.aca.2018.05.015
0003-2670/

70 A. Alibakshi / Analytica Chimica Acta 1026 (2018) 69–76

Flash point
Regression
Model developement

Based on the introduced considerations, an ANN model was proposed for predicting the flash point of pure organic compounds. According to the results, the new model was found to produce the lowest error compared to other available models.

1. Introduction

Artificial neural network (ANN) is one of the most efficient tools that work based on artificial intelligence and machine learning. ANNs are capable of doing several tasks such as function approximation [1], pattern recognition [2], data clustering [3], prediction of time series [4], and so on. To provide the best performance, various types of neural networks are developed and characterized depending on the application. However, despite their slight differences, all of them follow the same basics taken from the learning mechanisms of the biological neural networks [5].

Model development which is a function approximation problem, is probably the most widely used application of the ANNs in chemistry and chemical engineering [6–11]. The most appropriate ANN for model development is the multilayer network shown in Fig. 1, known as the feedforward neural network. For model development using a multilayer feedforward neural network, the input variables are introduced to the network as a vector and are processed by the neurons of the first layer. Each neuron in the first layer is connected to all of the input variables and for each connection, a weight constant is assigned. The summation of all input variables multiplied by their respective weights and a bias constant yields the input of each neuron. A transfer function modifies the inputs to result the output of each neuron which is then transmitted to the neurons of the next layer to be processed further in the same way.

To develop an ANN model, the weights, biases and transfer functions are determined in a way that each set of inputs result in a final output equivalent to the required property. To do so, using a dataset with known input and output data (training dataset), the optimum values for the constants are determined through a procedure which is called the training of the network. Various training algorithms are developed and can be used for this purpose. With appropriate number of neurons, transfer functions and training algorithm, a multilayer feedforward ANN is capable of modeling any linear or non-linear correlation between the input and output variables [12].

Two widely used methods which apply ANNs to model properties of chemical compounds are group contribution method (GCM) and quantitative structure property relationship (QSPR).

According to the GCM, properties of a compound are predicted based on the number and types of its constituting functional groups. The simplest form of the GCM is written as:

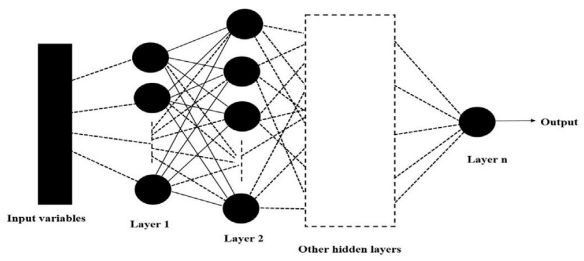


Fig. 1. The configuration of a feedforward neural network.

$$\varphi = c + \sum n_i \varphi_i \quad (1)$$

where φ is the required property, n_i and φ_i are the number of presence and amount of contribution of functional group i, respectively, and c is a constant.

Prediction of properties via equation (1) is known as the Joback method. The Joback method typically produces poor results for large datasets. However, these results become considerably more accurate when the correlation between functional groups and the required property are mapped via ANNs. For example, using a feedforward neural network with one hidden layer containing 7 neurons, Albahri could predict the flash points of 375 transportation fuels based on the GCM with an average absolute relative error (*AARE*) of 1.1%, while the Joback method resulted an *AARE* of 4.3% for the same dataset and functional groups [13].

The group contribution based models which use ANNs have been widely used to predict various properties such as liquid viscosity [14,15], thermal conductivity [16], infinite dilution activity [17], and density of ionic liquids [18], normal boiling point (NBP) [19], flash point (FP) [13] and melting point [20].

Contrary to the classic GCM which only considers the functional groups as contributors to a property, the QSPR applies a more extensive set of structure based quantities, known as molecular descriptors, to model a property. To develop a QSPR model, the most effective molecular descriptors are screened from a pool of numerous calculated descriptors and are used as the inputs of the model. ANN based QSPR models have also been extensively used to predict various properties e.g. NBP [21], FP [22–24], surface tension [25], ideal gas entropy [26], aqueous solubility [27], Hildebrand solubility parameter [28] and so on.

Using several available software tools e.g. Matlab, R, and Neurosolutions, developing an ANN model has become considerably straightforward, without requiring any knowledge of its extensive theoretical details. While simplified in practice, ANN models soon become unreliable without full consideration of important details e.g. appropriate assignment of the number of neurons, number of layers, and training of the network. The present study discusses such details and introduces the practical aspects of developing a robust ANN model which can be used for other models developed based on supervised learning. As a case study, predicting the flash point (FP) for an extensive dataset via a two layer feedforward neural network is investigated. The FP is one of the most important flammability properties of chemical compounds in assessment of fire hazards [29], and its predictability via ANNs has been widely studied in many works [13,30–32].

2. Practical aspects of developing ANN models

2.1. Dataset division

Similar to any model developed based on supervised learning, the first step in developing an ANN model is to divide the dataset into three subsets, namely training, validation and test datasets. The training dataset is used to train the model, where an error

function which is usually the average absolute relative error or mean squared error is minimized with respect to the model constants in successive iterations. The number of compounds needed for training, as discussed later, is the first important factor affecting the reliability of an ANN model and determines the number of neurons and layers.

As the training goes on, the performance of the model is continuously improved for the training dataset, which simultaneously increases the risk of overfitting as well. Overfitting causes a model to yield accurate results for the dataset used for developing that model but poor results for the compounds out of this dataset. To prevent overfitting, the performance of the studied ANN is simultaneously monitored and validated for an independent dataset which is called the validation dataset. An increase in the error function of the validation dataset in several successive iterations is an indicator of overfitting and is used as a condition to stop the training. Once a model is trained, i.e. the optimum values of all constants are determined, the performance of the model is examined using another independent dataset, known as the test dataset. Usually, 60–80% of the dataset is assigned for training, 10–20% for validation and 10–20% to test the model.

2.2. Assigning the number of neurons and hidden layers

After specifying the training, validation and test datasets, selecting the number of layers, number of neurons in each layer, type of transfer functions and training algorithm and assigning the initial values for the weights and biases are the subsequent steps to develop an ANN model.

The number of layers and neurons in each layer is one of the most crucial parameters affecting the performance and reliability of an ANN model. With a higher number of layers or neurons, an ANN typically yields more accurate results and can model more complicated relationships. However, it can also highly increase the risk of overfitting, simultaneously.

There are some recommendations to evaluate the appropriate number of hidden layer neurons, e.g. setting a number of hidden layer neurons equal to 2/3 of the number of input layer neurons [33], between the number of neurons in the input and output layers [34], or lower than twice the number of neurons in the input layer [35]. However, such recommendations don't seem to be very robust as they totally neglect the number of training samples and details of the ANN configuration as the most important factors.

Considering an ANN model as a regression problem, the ratio of the training samples to the total number of ANN constants as suggested by Jackson for regression models [36], can be used as an index for determining the appropriate number of neurons and layers. For a multilayer feedforward neural network, if the number of input variables is V_n, the number of neurons in the hidden layer i is N_i, and the number of layers is n_L, the total number of weight and bias constants (S) can be calculated via:

$$S = V_n N_1 + \sum_{i=1}^{n_L-1} N_i N_{i+1} + \sum_{i=1}^{n_L} N_i. \quad (2)$$

Obviously, the higher the ratio of the training samples to S, the higher the reliability of the results obtained by the model. It will be further discussed in section 4.2.

2.3. Assigning training algorithm and transfer function

The applied transfer functions and training algorithm are also other ANN parameters which can have impacts on the performance of an ANN [37,38]. Some commonly used transfer functions in ANN models which are also considered in our study are hyperbolic tangent sigmoid (*tansig*), log-sigmoid (*logsig*), Hard-limit (*hardlim*), Positive linear (*poslin*), and Radial basis (*radbas*) transfer functions for the hidden layers, and linear transfer function (*purelin*) for the output layer. The most widely used training algorithms are reported in Table 1.

2.4. Initialization of the network

Assignment of the initial values for the weight and bias constants (initialization) is the last step before we can start training of the network. Initialization is a necessary task as the efficient training algorithms, all require some initial values for the weights and biases to optimize them in successive iterations via minimizing the error function. The error function which should be minimized with respect to the weight and bias constants, typically has several local minima. As a result, starting from different initialization states determined by the initial values of the weights and biases, we may get to a quite different local minimum and obtain considerably different results. Some theoretical approaches on appropriate initialization of an ANN is reviewed by Yam and Chow [40]. An easy and yet very efficient approach to overcome the problems originated by initialization is to repeat the training of the network for various initialization states [41], which will be taken in this work to show the effect and importance of initialization and also to overcome the discussed problems.

2.5. Evaluating the reliable performance of an ANN model

As discussed before and will be seen in the results, the performance of an ANN highly depends on the ANN specifications (applied dataset division, transfer function, training algorithm, initialization state and so on). Now the crucially important questions that arise are which ANN provides the most reliable model and among all different results obtained for each ANN configuration using different initialization or dataset division states, which one represent the authentic performance of that configuration? One may argue that the ANN which provides the most accurate results for the overall or test dataset is the one to choose, however, such data won't be reliable. An excellent performance for the

Table 1
Appropriate training algorithms in model development [39].

Training algorithm	Abbreviation
Levenberg-Marquardt backpropagation	*trainlm*
Gradient descent backpropagation	*traingd*
Resilient backpropagation	*trainrp*
Scaled conjugate gradient backpropagation	*trainscg*
BFGS quasi-Newton backpropagation	*trainbfg*
Conjugate gradient backpropagation with Fletcher-Reeves updates	*traincgf*
Gradient descent with momentum backpropagation	*traingdm*

overall dataset may be affected by overfitting and for the test dataset it can be just the result of a lucky dataset division. Another option is to use the average of results obtained for all studied initialization and dataset division states for an ANN. However, average of all results is not always informative because as discussed before, due to the high number of constants in many ANN models, error function typically has several local minima and therefore, initialization plays an important role. As a result, it is quite plausible that only a few initialization states may result in an accurate and reliable model (see e.g. the results obtained for developed FP predictive ANNs with 4 neurons, reported in the supplementary material) and therefore, the average of all results won't represent the authentic performance of an ANN in most cases. In other words, a good initialization state may get lost among several other inappropriate initialization states by averaging.

To overcome this issue, in the present study instead of averaging the results of all initialization and dataset division states, for each individual initialization state we retrain the model for several different dataset division states and use their average as the authentic performance of that configuration and parameters. To find the most reliable model, the ANN for which in most of the repeats the observed errors of the training and test sets are not significantly different confirmed by suitable statistical tests, can be considered as the appropriately trained and reliable models (low risk of overfitting).

3. Material and method

3.1. Dataset

To develop a robust model, reliability of the dataset used for model development plays an important role. In the present study, the DIPPR 801 [42] database was used to evaluate the predictability of the FP for an extensive dataset of pure organic compounds. DIPPR 801 provides evaluated data for several properties is one of the most widely used databases to develop FP predictive models.

To implement a GCM based model, number of presence of the functional groups listed in Table 2 and the experimentally determined data of the NBP were used as the inputs of the model.

Normal boiling point and enthalpy of vaporization have been used is many FP predictive models [43–46], as both of them represent the volatility and hence, flammability of a fuel [47]. In previous studies, it was shown that considering a contribution for the NBP in addition to the same functional groups listed in Table 2, can significantly improve the predictability of the FP [43,48]. Among the organic compounds available in the DIPPR database, 740 compounds were available for which the reported data for both FP and NBP were experimentally determined. For other compounds, as a predicted data were reported for at least one of those properties, they were not considered in model development and evaluation. The full list of studied compounds can be found as supplementary material. The NBP was assigned to the first neuron in the input layer and the number of presence of the functional groups 1 to 42 were processed by the neurons 2–43, respectively.

3.2. Initial implementation of the ANNs

To study the various parameters affecting the performance of an ANN, the predictability of the FP for the dataset of 740 pure organic compounds from diverse families was initially investigated using feedforward neural networks with 1–10 neurons in the hidden layer. 75% of the dataset was assigned for training, 13% for validation, and 12% to test the ANN models.

To study the impacts of dataset division on the performance of ANNs, randomly division of the dataset was repeated 20 times. To investigate the impacts of various training algorithms, transfer functions and initialization states, for each dataset division the *trainlm, traingd, trainrp, trainscg, trainbfg, traincgf*, and *traingdm* training algorithms and *tansig, logsig, hardlim, poslin,* and *radbas* transfer functions were examined for 20 different initialization states. An increase in the mean squared error of the validation dataset in 6 successive iterations was considered as the condition to stop the training.

Therefore, considering 20 different dataset division states and for each one, assigning 1 to 10 neurons for the hidden layer, 5 different transfer functions, 7 different training algorithms, and 20 different initialization states, a total number of 140000 neural networks were initially implemented and their performance for FP prediction were evaluated using a Matlab code. The interested readers can get the Matlab code by contacting the corresponding author. The performance of the ANNs were reported as percentage average absolute relative errors (*AARE%*) and correlation coefficients (R) defined as:

$$AAD = \frac{1}{N}\sum\left(\left|y_i^{ref} - y_i^{pred}\right|\right), \quad (3)$$

Table 2
The constituting functional groups used in group contribution method.

	Functional Group
1	−CH_3
2	−CH_2−
3	>CH−
4	>C<
5	=CH_2
6	=CH−
7	=C<
8	=C=
9	≡CH
10	≡C−
11	−OH
12	−O−
13	>C=O
14	−CHO (aldehyde)
15	−COOH (acid)
16	−COO− (ester)
17	HCOO− (formate)
18	−NH_2
19	−NH−
20	>N−
21	=N−
22	−C≡N
23	−NO_2
24	−F
25	−Cl
26	−Br
27	−I
28	−SH
29	−S−
30	−CH_2− (ring)
31	−HC< (ring)
32	=CH− (ring)
33	>C< (ring)
34	=C< (ring)
35	−O− (ring)
36	−OH (ring)
37	>C=O (ring)
38	−NH− (ring)
39	>N− (ring)
40	=N− (ring)
41	−S− (ring)
42	-CO-O-CO- (anidride)

$$AARE\% = \frac{1}{N}\sum\left(\left|\frac{y_i^{exp} - y_i^{pred}}{y_i^{exp}}\right|\right) \times 100, \quad (4)$$

$$R = \frac{N\sum y_i^{exp} y_i^{pred} - \sum y_i^{exp} \sum y_i^{pred}}{\sqrt{N\sum\left(y_i^{exp}\right)^2 - \left(\sum y_i^{exp}\right)^2}\sqrt{N\sum\left(y_i^{pred}\right)^2 - \left(\sum y_i^{pred}\right)^2}}, \quad (5)$$

where y_i^{exp} and y_i^{pred} are the experimentally determined and predicted values of FP, respectively.

For the ANNs which resulted an overall *AARE%* of lower than 1.5%, the initially assigned constants of the network were recorded to be used for evaluation of their authentic performances as explained in section 2.5 in the next step.

3.3. Evaluation of the authentic performance of ANNs

After initially developing 140000 ANN models in the previous step, the models which resulted an *AARE%* of less than 1.5% were selected for retraining and evaluating their authentic performance, based on the approach discussed in section 2.5. To do so, the initially assigned weight and bias constants recorded in the previous step were used to retrain the selected models for 20 different random dataset division states. The average of the results for 20 repeats were considered as the reliable performance for each configuration and initialization state. The two sample *t*-test method was used to compare the results obtained for the training and test datasets. The models for which the relative errors of the training and test datasets were not significantly different with 95% of significance level in at least 19 repeats, were considered as reliable and efficiently trained models.

4. Results and discussion

4.1. Implementation of ANNs

The effect of various parameters on the performance of an ANN can be observed in the results of 140000 initially developed ANN models reported in the supplementary material. A quick overview of the initially obtained results shows that for each dataset division state changing one of the studied parameters i.e. the initialization state, number of hidden layer neurons, training algorithm or transfer function while the other parameters remain unchanged can yield considerably different results.

Among all initially studied ANNs, 17211 models yielded an overall *AARE%* of lower than 1.5% and retrained for 20 different dataset division states to evaluate their authentic performance, as discussed before. For each model, the average of 20 repeat results were calculated and are used in the next sections to evaluate the effect of each parameters. The details of the repeated results for each configuration are reported in the supplementary material.

4.2. The effect of assigned number of hidden layer neurons

As the first important factor, we consider the effect of the number of hidden layer neurons. To do so, we exploited the percentage of the initially selected ANNs with overall *AARE%* of lower than 1.5% which after retraining for 20 different dataset division states and averaging, resulted an overall *AARE%* of lower than 1.5% again. Based on the results reported in Table 3, a remarkable difference between the ANNs with one neuron in the hidden layer and other ANNs with higher number of neurons can be clearly observed. According to the results, 99.91% of the ANNs with 1 neuron in the hidden layer which yielded an *AARE%* of lower than 1.5% in the first step, after retraining and averaging yielded an *AARE %* of lower than 1.5% again, while for other ANNs with higher number of neurons this value was lower than 42%. The same remarkable difference can also be observed for the average of standard deviations of *AARE%* calculated for the 20 repeats of each configuration, reported in Table 3.

Using the average results of 20 repeats and applying the *t*-test method, for each number of hidden layer neurons the percentage of reliable models (the models for which in at least 19 repeats the errors of the training and test sets were with 95% of significance level the same) were calculated and are depicted in Fig. 2.

According to the results depicted in Fig. 2, 27.66% of ANNs with 1 neuron in the hidden layer were efficiently trained, while this value for other ANNs with higher number of neurons is lower than 4.3%.

Based on the abovementioned observations, we can conclude that for our case study the optimum number of hidden layer neurons is one. Considering the number of weight and bias constants for our case study, for ANNs with one neuron in the hidden layer the ratio of the number of training samples to the ANN constants calculated using equation (2) is 12.07. This value for ANNs with 2 neurons in hidden layer is 6.1 and for higher number of neurons reduces further.

The observed ratio of the training samples to the number of constants for ANNs with one hidden layer neuron is consistent with the value recommended by Kline who suggested a value of atleast 10 for this ratio to avoid overfitting in regression models [49].

Therefore for developing reliable ANN models, considering the number of training samples and ANN constants, existence of at least 10 samples in the training set for each constant can be used as a rule of thumb. It means, considering the number of training

Table 3
The percentage of initially implemented ANNs which after retraining yielded *AARE%* of lower than 1.5% again and the average of standards deviations obtained for 20 repeat of each configuration.

Nr. of hidden layer neurons	Ratio of the ANNs which reproduce the initially obtained results	Mean of average standard deviations in 20 repeats (overall)	Mean of average standard deviations in 20 repeats (training)	Mean of average standard deviations in 20 repeats (validation)	Mean of average standard deviations in 20 repeats (test)
1	99.91	0.015289	0.039167	0.13143	0.15645
2	41.35	1.1976	1.2049	1.286	1.3726
3	40.74	0.99861	1.0024	1.0813	1.1486
4	40.3	0.90107	0.90617	0.96768	1.0179
5	42.79	0.82164	0.82722	0.87261	0.91792
6	42.92	0.80169	0.80691	0.8547	0.90352
7	42.93	0.78356	0.79009	0.82401	0.86634
8	44.69	0.73719	0.74234	0.7822	0.82574
9	46.85	0.78592	0.79338	0.82743	0.85918
10	46.66	0.76154	0.76925	0.80203	0.83607

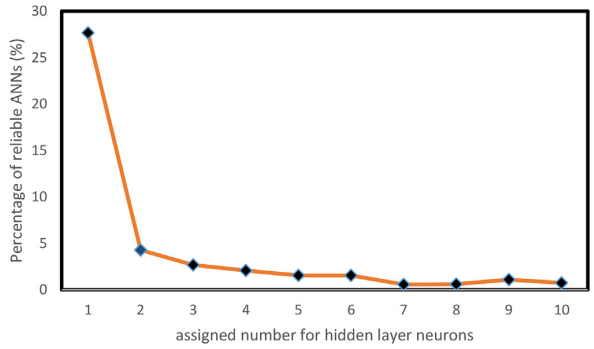


Fig. 2. The percentage of reliable ANNs for each assigned number of hidden layer neurons.

samples, the number of layers and neurons should be selected in a way that the number of constants calculated via equation (2) is one tenth or smaller than the number of training samples. A literature survey shows many models which considerably deviate from this condition. Some of those models are reported in Table 4.

4.3. Impact of training algorithm and transfer function

The impact of the training algorithm and transfer function was investigated via analyzing the performance of efficiently trained models in FP prediction for each combination of training algorithm and transfer function. The results are reported in Table 5 and show that different combinations of training algorithms and transfer functions can yield quite different results. More ever, for prediction of FP applying the *trainlm* training algorithm and *logsig* transfer function seems to be more efficient.

4.4. The selected ANN for FP prediction

Among the efficiently trained models, the best performance was observed for an ANN model with 1 neuron in the hidden layer, *trainlm* training algorithm and *logsig* transfer function. For this configuration the average of 20 repeats resulted an *AARE%* of 1.198, 1.194, 1.195 and 1.122 and correlation coefficients of 0.9933, 0.9933, 0.9933 and 0.9934, for the overall, training, validation and test datasets, respectively. More ever, the relative errors of the training and test datasets in none of the 20 repeats were significantly different. For this ANN, the initial values of the weight and bias constants are reported in supplementary material.

The obtained results are compared with those of the most accurate ever proposed models in Table 6. As can be seen in Table 6, only one model provides better results than the current model which was proposed by Albahri [13]. However, the Albahri's model is developed for a smaller set of chemicals which are transportation fuels and the diversity of the chemical families is lower than the dataset studied in the present work. Furthermore, as discussed before for the Albahri's model the ratio of training samples to constants is 1.22 which implies a high risk of overfitting in that model.

To remind the importance of the approach used to evaluate the authentic performance of ANNs, it should be noted that among the initially developed models, an overall *AARE%* as low as 0.72% was also observed for an ANN with 10 neurons in the hidden layer which is surely resulted by overfitting.

Table 4
Examples of lack of sufficient data in the training dataset in ANN based FP predictive models.

	Number of weight and bias constants	Number of compounds in the training dataset	Ratio of the training samples to constants
Gharagheizi [50]	1481	821	0.55
Lazzús [51]	353	328	0.93
Lazzús [30]	369	350	0.95
Valderrama et al. [18],	421	399	0.95
Albahri [13]	274	335	1.22
Fathi et al. [52].	316	550	1.53
Gharagheizi [53]	536	846	1.58

Table 5
Details of reliable ANNs for each combination of transfer function and training algorithm.

Training algorithm	Transfer function	Nr. reliable models	*AARE%* (overal)	*AARE %* (training)	*AARE %* (validation)	*AARE %* (test)	Training algorithm	Transfer function	Nr. reliable models	*AARE %* (overal)	*AARE %* (training)	*AARE %* (validation)	*AARE %* (test)
trainlm	logsig	70	1.57	1.57	1.57	1.58	*trainbfg*	logsig	51	2.08	2.06	2.12	2.12
trainlm	tansig	56	1.88	1.87	1.89	1.89	*trainbfg*	tansig	3	8.74	8.74	8.6	8.88
trainlm	hardlim	0					*trainbfg*	hardlim	0				
trainlm	poslin	73	1.61	1.61	1.62	1.61	*trainbfg*	poslin	46	8.01	8	8.02	8.05
trainlm	radbas	25	5.09	5.07	5.15	5.18	*trainbfg*	radbas	1	5.61	5.61	5.61	5.61
traingd	logsig	0					*traincgf*	logsig	23	4.86	4.86	4.78	4.92
traingd	tansig	0					*traincgf*	tansig	16	4.18	4.18	4.16	4.24
traingd	hardlim	0					traincgf	hardlim	0				
traingd	poslin	0					*traincgf*	poslin	80	4.84	4.84	4.81	4.88
traingd	radbas	0					traincgf	radbas	5	6.34	6.34	6.24	6.48
trainrp	logsig	6	3.58	3.57	3.62	3.59	traingdm	logsig	0				
trainrp	tansig	0					*traingdm*	tansig	0				
trainrp	hardlim	0					*traingdm*	hardlim	0				
trainrp	poslin	12	6.94	6.95	6.89	6.95	traingdm	poslin	0				
trainrp	radbas	0					traingdm	radbas	0				
trainscg	logsig	23	1.76	1.75	1.78	1.81							
trainscg	tansig	13	2.94	2.94	2.93	2.94							
trainscg	hardlim	0											
trainscg	poslin	44	3.08	3.07	3.11	3.12							
trainscg	radbas	1	1.47	1.47	1.47	1.47							

Table 6
Comparison of the results of the developed model with other accurate models reported in the literature.

Model	Method	No. data	AAD (k)	*AARE (%)*	Max. *AARE* (%)	R
The new model (overall)	GCM + ANN	740	–	1.198	–	0.9933
Alibakhshi et.al. [43]	Semi- empirical	740	4.066	1.225	9.81	0.9934
Alibakhshi et. al. [48]		740	4.11	1.23	9.49	0.9935
Albahri [13]	GCM + ANN	375	3.55	1.1	6.62	0.9961
Rowley et al. [54]	Correlation (ΔH_v+NBP)	1062	4.65	1.32	–	–
Lazzús [30]	GCM + ANN + PSO	505	6.2	1.8	8.6	–
Catoire &Naudet [55]	Correlation (ΔH_v+NBP)	600	6.36	1.84	–	–
Mathieu [56]	Correlation	92	3.75	1.37	5.4	0.9922
Pan et al. [57]		92	3.75	1.38	10.18	0.9907
Keshavarz and Ghanbarzadeh [58]	Correlation	173	6.35	2.21	12.8	0.9899
Mathieu and Alaime [59]	–	488	8.6	–	–	–
Rowley et al. [60]	Correlation (ΔH_v+NBP)	1062	9.68	2.84	–	–
Tetteh et al. [32]	QSPR + ANN	400	9.59	–	–	–
Hukkerikar et al. [61]	GC^+	512	10.66	3.27	–	0.89
Mathieu [44]	QSPR	230	12	–	–	0.943
Keshavarz et al. [62]	Correlation	548	12.1	–	–	–
Katritzky et al. [31]	QSPR + ANN	758	12.6	–	–	0.989
Khaje and Modarres [63]	ANFIS	95	11.5	31.1	1500	0.986
Khaje and Modarres [63]	GFA	95	13.08	25.8	966.75	0.98
Chen et. al. [64]	QSPR	230	–	–	22.9	0.964
Hshieh [65]	Correlation (NBP)	494	–	–	–	0.966
Bagheri et al. Bagheri, 2012 #33}	QSPR	1651	19.31	5.94	–	0.94
Katritzky et al. [66]	QSPR	271	–	–	–	0.91
Patil [67]	Correlation (NBP)	593	–	–	7.5	0.90

Among initially developed ANNs, we can also find a model with 9 neurons in the hidden layer for which the *AARE%* of the test and training dataset were not significantly different and yielded an overall *AARE%* as low as 0.84%. However after retraining this model for same initial parameters and different dataset division states, we find that this model wouldn't reproduce such excellent results anymore, which implies that the initially obtained results is affected by overfitting. This confirms the importance of the proposed approach for evaluating the authentic performance of each model.

5. Conclusion

In the present study we studied various parameters which can affect the reliability and performance of ANN models. Among all affecting parameters, appropriate selection of the number of neurons and hidden layers seems to be the highest priority which should be determined based on the number of training samples. The results show that considering the ratio of the training samples to ANN constants can be used as an index for evaluating the appropriate number of neurons and layers. Our results suggest a value of equal or greater than 10 for this ratio which has already been suggested for regression models elsewhere too. The second crucially important strategy in developing reliable ANN models is to study various initialization states for the ANNs and for each one, repeating the randomly division of the dataset and use the average of those repeats as the authentic performance obtained for each initialization state. Comparing the relative errors of the training and test datasets in different repeats using appropriate statistical tests can be used to find the efficiently trained models.

Considering various combinations of training algorithm and transfer functions to find the best combination of those parameters is also another efficient strategy to find more efficient ANN models.

Acknowledgment

The author sincerely thanks an anonymous reviewer whose careful reading and insightful comments caused a substantial improvement of the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2018.05.015.

References

[1] B. Tripp, C. Eliasmith, Function approximation in inhibitory networks, Neural Network. 77 (2016) 95–106.
[2] M. Alfaro-Ponce, A. Argüelles, I. Chairez, Pattern recognition for electroencephalographic signals based on continuous neural networks, Neural Network. 79 (2016) 88–96.
[3] C. Hajjar, H. Hamdan, Interval data clustering using self-organizing maps based on adaptive Mahalanobis distances, Neural Network. 46 (2013) 124–132.
[4] Y. Cui, J. Shi, Z. Wang, Complex rotation quantum dynamic neural networks (CRQDNN) using complex quantum neuron (CQN): applications to time series prediction, Neural Network. 71 (2015) 11–26.
[5] I.N. da Silva, D.H. Spatti, R.A. Flauzino, L.H.B. Liboni, S.F. dos Reis Alves, Artificial Neural Networks: a Practical Course, Springer, 2016.
[6] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T.N. Tran, L.M. Buydens, E. Marchiori, Convolutional neural networks for vibrational spectroscopic data analysis, Anal. Chim. Acta 954 (2017) 22–31.
[7] I. Kuzmanovski, A. Wagner, M. Novič, Development of models for prediction of the antioxidant activity of derivatives of natural compounds, Anal. Chim. Acta 868 (2015) 23–35.
[8] N. Fjodorova, M. Novič, Searching for optimal setting conditions in technological processes using parametric estimation models and neural network mapping approach: a tutorial, Anal. Chim. Acta 891 (2015) 90–100.
[9] W. Ni, L. Nørgaard, M. Mørup, Non-linear calibration models for near infrared spectroscopy, Anal. Chim. Acta 813 (2014) 1–14.
[10] Q. Ouyang, J. Zhao, Q. Chen, Instrumental intelligent test of food sensory quality as mimic of human panel test combining multiple cross-perception sensors and data fusion, Anal. Chim. Acta 841 (2014) 68–76.
[11] N. Minovski, Š. Župerl, V. Drgan, M. Novič, Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: a case study, Anal. Chim. Acta 759 (2013) 28–42.
[12] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (5) (1989) 359–366.
[13] T.A. Albahri, MNLR and ANN structural group contribution methods for predicting the flash point temperature of pure compounds in the transportation fuels range, Process Saf. Environ. Protect. 93 (2015) 182–191.
[14] T. Suzuki, R.-U. Ebert, G. Schüürmann, Application of neural networks to modeling and estimating temperature-dependent liquid viscosity of organic compounds, J. Chem. Inf. Comput. Sci. 41 (3) (2001) 776–790.
[15] K. Paduszynski, U. Domanska, Viscosity of ionic liquids: an extensive database and a new group contribution model based on a feed-forward artificial neural network, J. Chem. Inf. Model. 54 (5) (2014) 1311–1324.

[16] A.Z. Hezave, S. Raeissi, M. Lashkarbolooki, Estimation of thermal conductivity of ionic liquids using a perceptron neural network, Ind. Eng. Chem. Res. 51 (29) (2012) 9886–9893.
[17] K. Paduszynski, In silico calculation of infinite dilution activity coefficients of molecular solutes in ionic liquids: critical review of current methods and new models based on three machine learning algorithms, J. Chem. Inf. Model. 56 (8) (2016) 1420–1437.
[18] J.O. Valderrama, A. Reátegui, R.E. Rojas, Density of ionic liquids using group contribution and artificial neural networks, Ind. Eng. Chem. Res. 48 (6) (2009) 3254–3259.
[19] J.A. Palatinus, C.M. Sams, C.M. Beeston, F.A. Carroll, A.B. Argenton, F.H. Quina, Kinney revisited: an improved group contribution method for the prediction of boiling points of acyclic alkanes, Ind. Eng. Chem. Res. 45 (20) (2006) 6860–6863.
[20] J.A. Lazzús, Hybrid method to predict melting points of organic compounds using group contribution+ neural network+ particle swarm algorithm, Ind. Eng. Chem. Res. 48 (18) (2009) 8760–8766.
[21] A.J. Chalk, B. Beck, T. Clark, A quantum mechanical/neural net model for boiling points with error estimation, J. Chem. Inf. Comput. Sci. 41 (2) (2001) 457–462.
[22] L. Jiao, X. Zhang, Y. Qin, X. Wang, H. Li, QSPR study on the flash point of organic binary mixtures by using electrotopological state index, Chemometr. Intell. Lab. Syst. 156 (15 August 2016) 211–216.
[23] L.Y. Phoon, A.A. Mustaffa, H. Hashim, R. Mat, A review of flash point prediction models for flammable liquid mixtures, Ind. Eng. Chem. Res. 53 (32) (2014) 12553–12565.
[24] S.J. Patel, D. Ng, M.S. Mannan, QSPR flash point prediction of solvents using topological indices for application in computer aided molecular design, Ind. Eng. Chem. Res. 48 (15) (2009) 7378–7387.
[25] T.A. Albahri, D.A. Alashwak, Modeling of pure compounds surface tension using QSPR, Fluid Phase Equil. 355 (2013) 87–91.
[26] A. Fazeli, M. Bagheri, S. Ghaniyari-Benis, R. Aslebagh, E. Kamaloo, Prediction of absolute entropy of ideal gas at 298K of pure chemicals through GAMLR and FFNN, Energy Convers. Manag. 52 (1) (2011) 630–634.
[27] L. Jiao, H. Li, QSPR studies on the aqueous solubility of PCDD/Fs by using artificial neural network combined with stepwise regression, Chemometr. Intell. Lab. Syst. 103 (2) (2010) 90–95.
[28] M. Goodarzi, P.R. Duchowicz, M.P. Freitas, F.M. Fernández, Prediction of the Hildebrand parameter of various solvents using linear and nonlinear approaches, Fluid Phase Equil. 293 (2) (2010) 130–136.
[29] A. Alibakhshi, H. Mirshahvalad, S. Alibakhshi, Investigating the mechanism of effect of carbon nanotubes on flame spread over liquid fuels, Fire Technol. 51 (4) (2015) 759–770.
[30] J.A. Lazzús, Prediction of flash point temperature of organic compounds using a hybrid method of group contribution+ neural network+ particle swarm optimization, Chin. J. Chem. Eng. 18 (5) (2010) 817–823.
[31] A.R. Katritzky, I.B. Stoyanova-Slavova, D.A. Dobchev, M. Karelson, QSPR modeling of flash points: an update, J. Mol. Graph. Model. 26 (2) (2007) 529–536.
[32] J. Tetteh, T. Suzuki, E. Metcalfe, S. Howells, Quantitative structure-property relationships for the estimation of boiling point and flash point using a radial basis function neural network, J. Chem. Inf. Comput. Sci. 39 (3) (1999) 491–507.
[33] Z. Boger, H. Guterman, In Knowledge extraction from artificial neural network models, Systems, Man, and Cybernetics, 1997, in: Computational Cybernetics and Simulation., 1997 IEEE International Conference on, IEEE, 1997, pp. 3030–3035.
[34] A. Blum, Neural Networks in C++: an Object-oriented Framework for Building Connectionist Systems, John Wiley & Sons, Inc, 1992.
[35] M.J. Berry, G. Linoff, Data Mining Techniques: for Marketing, Sales, and Customer Support, John Wiley & Sons, Inc, 1997.
[36] D.L. Jackson, Revisiting sample size and number of parameter estimates: some support for the N: q hypothesis, Struct. Equ. Model. 10 (1) (2003) 128–141.
[37] H. Yonaba, F. Anctil, V. Fortin, Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting, J. Hydrol. Eng. 15 (4) (2010) 275–283.
[38] B. Karlik, A.V. Olgac, Performance analysis of various activation functions in generalized MLP architectures of neural networks, Int. J. Artif. Intell. Expet. Syst. 1 (4) (2011) 111–122.
[39] H. Demuth, M. Beale, Neural Network Toolbox for Use with MATLAB, 1993.
[40] J.Y. Yam, T.W. Chow, A weight initialization method for improving training speed in feedforward neural network, Neurocomputing 30 (1) (2000) 219–232.
[41] C. MacLeod, An Introduction to Practical Neural Networks and Genetic Algorithms for Engineers and Scientists, Jerhi Wahyu Fernanda, 2013, p. 157.
[42] American Institute of Chemical Engineers (AIChE), E P D d, public release, documentation, design institute for physical properties (DIPPR), Project 801, 2016.
[43] A. Alibakhshi, H. Mirshahvalad, S. Alibakhshi, Prediction of flash points of pure organic compounds: evaluation of the DIPPR database, Process Saf. Environ. Protect. 105 (2017) 127–133.
[44] D. Mathieu, Flash points of organosilicon compounds: how data for alkanes combined with custom additive fragments can expedite the development of predictive models, Ind. Eng. Chem. Res. 51 (43) (2012) 14309–14315.
[45] F.A. Carroll, C.-Y. Lin, F.H. Quina, Simple method to evaluate and to predict flash points of organic compounds, Ind. Eng. Chem. Res. 50 (8) (2011) 4796–4800.
[46] M. Riazi, T. Daubert, Predicting flash and pour points, Hydrocarb. Process. 66 (9) (1987) 81–83.
[47] A. Alibakhshi, Enthalpy of vaporization, its temperature dependence and correlation with surface tension: a theoretical approach, Fluid Phase Equil. 432 (2017) 62–69.
[48] A. Alibakhshi, H. Mirshahvaland, S. Alibakhshi, A modified group contribution method for accurate prediction of flash points of pure organic compounds, Ind. Eng. Chem. Res. 54 (44) (2015) 11230–11235.
[49] R.B. Kline, Principles and Practice of Structural Equation Modeling, Guilford publications, 2015.
[50] F. Gharagheizi, An accurate model for prediction of autoignition temperature of pure compounds, J. Hazard Mater. 189 (1) (2011) 211–221.
[51] J.A. Lazzús, Prediction of flammability limit temperatures from molecular structures using a neural network–particle swarm algorithm, Journal of the Taiwan Institute of Chemical Engineers 42 (3) (2011) 447–453.
[52] M.-R. Fatehi, S. Raeissi, D. Mowla, Estimation of viscosity of binary mixtures of ionic liquids and solvents using an artificial neural network based on the structure groups of the ionic liquid, Fluid Phase Equil. 364 (2014) 88–94.
[53] F. Gharagheizi, A new group contribution-based model for estimation of lower flammability limit of pure compounds, J. Hazard Mater. 170 (2) (2009) 595–604.
[54] J.R. Rowley, R.L. Rowley, W.V. Wilding, Prediction of pure-component flash points for organic compounds, Fire Mater. 35 (6) (2011) 343–351.
[55] L. Catoire, V. Naudet, A unique equation to estimate flash points of selected pure liquids application to the correction of probably erroneous flash point values, J. Phys. Chem. Ref. Data 33 (4) (2004) 1083–1111.
[56] D. Mathieu, Inductive modeling of physico-chemical properties: flash point of alkanes, J. Hazard Mater. 179 (1) (2010) 1161–1164.
[57] Y. Pan, J. Jiang, Z. Wang, Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network, J. Hazard Mater. 147 (1) (2007) 424–430.
[58] M.H. Keshavarz, M. Ghanbarzadeh, Simple method for reliable predicting flash points of unsaturated hydrocarbons, J. Hazard Mater. 193 (2011) 335–341.
[59] D. Mathieu, T. Alaime, Insight into the contribution of individual functional groups to the flash point of organic compounds, J. Hazard Mater. 267 (2014) 169–174.
[60] J. Rowley, R. Rowley, W. Wilding, Estimation of the flash point of pure organic chemicals from structural contributions, Process Saf. Prog. 29 (4) (2010) 353–358.
[61] A.S. Hukkerikar, S. Kalakul, B. Sarup, D.M. Young, G. r Sin, R. Gani, Estimation of environment-related properties of chemicals for design of sustainable processes: development of group-contribution+ (GC+) property models and uncertainty analysis, J. Chem. Inf. Model. 52 (11) (2012) 2823–2839.
[62] M.H. Keshavarz, M. Jafari, M. Kamalvand, A. Karami, Z. Keshavarz, A. Zamani, S. Rajaee, A simple and reliable method for prediction of flash point of alcohols based on their elemental composition and structural parameters, Process Saf. Environ. Protect. 102 (2016) 1–8.
[63] A. Khajeh, H. Modarress, QSPR prediction of flash point of esters by means of GFA and ANFIS, J. Hazard Mater. 179 (1) (2010) 715–720.
[64] C.-C. Chen, H.-J. Liaw, Y.-J. Tsai, Prediction of flash point of organosilicon compounds using quantitative structure property relationship approach, Ind. Eng. Chem. Res. 49 (24) (2010) 12702–12708.
[65] F.Y. Hshieh, Correlation of closed-cup flash points with normal boiling points for silicone and general organic compounds, Fire Mater. 21 (6) (1997) 277–282.
[66] A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, QSPR analysis of flash points, J. Chem. Inf. Comput. Sci. 41 (6) (2001) 1521–1530.
[67] G. Patil, Estimation of flash point, Fire Mater. 12 (3) (1988) 127–131.

Chapter 7

Practical models for evaluation of vapor-liquid phase change thermodynamics

Project overview and motivation:

Accurate evaluation of vapor-liquid phase change thermodynamics is of key importance in the majority of chemical industries. In chapter 5, we introduced a theoretical framework that can be employed for understanding the dependency of vaporization enthalpy to temperature and showed its efficiency in estimating other thermodynamic quantities associated with phase-change. This current project was an extension of the method introduced in chapter 5. We developed methods that allowed practical estimation of thermodynamically effective molecular surfaces using readily accessible thermophysical quantities and thus, allowing straightforward prediction of vaporization enthalpy as a function of temperature via correlations. Additionally, we also developed and provided machine learning models which can more rigorously employ the same model inputs required by the proposed correlations to achieve significantly higher accuracies.

Novelty aspects:

- A theoretically derived analytical expression proposed for accurate estimation of the vaporization enthalpy from melting point to the critical temperature by normal boiling point, critical temperature, and critical pressure as the only input variables. This correlation is significantly more advantageous compared to the previously reported

models which require either vaporization enthalpy at a reference temperature or acentric factor which requires measurement of saturation pressure for multiple temperatures

- The results which are obtained via machine learning models are currently the most accurate ever reported results for prediction of vaporization enthalpy at wide temperature ranges

Connection to other chapters:

The general idea behind the newly proposed methods in this study is based on the methods developed in chapter 5. The development of the machine-learning models is based on the guidelines introduced in chapter 6. One of the main inputs which remarkably improved the predictability of vaporization enthalpy in machine learning models was the cavity surfaces computed within the implicit solvent approaches and were inspired based on the knowledge acquired through studying those surfaces in chapter 4, 5, 8, and 9.

Contributions:

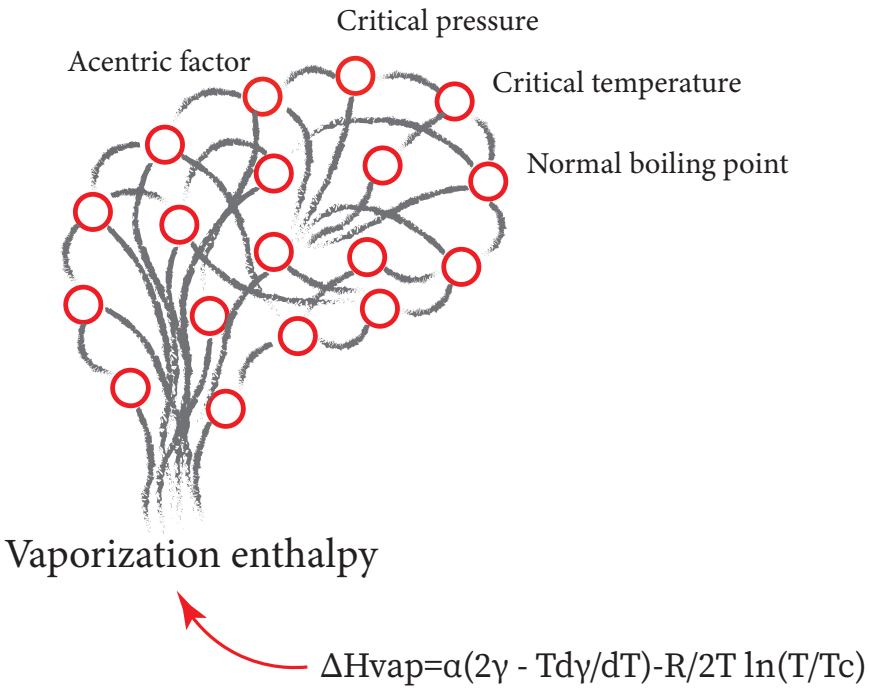
Major contribution in method development, carrying out all the computations, major contribution in writing the manuscript.

Publication status:

This work is in the submission process and its preprint is available online (DOI:10.26434/chemrxiv.13734007.v3).

Efficient models for high-accuracy evaluation of thermodynamic quantities associated with vapor-liquid phase change

Amin Alibakhshi[1,*], Bernd Hartke[1]
Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, 24118 Kiel, Germany
Corresponding author: alibakhshi@pctc.uni-kiel.de



Abstract

Evaluation of thermodynamic quantities associated with vapor-liquid phase change is extensively required in a very broad range of scientific fields. A commonly employed thermodynamic quantity that can characterize the vapor-liquid phase change thermodynamics is the vaporization enthalpy and its temperature dependence. In the present study, we introduced highly accurate and practical models for reliable prediction of vaporization enthalpy of compounds from diverse chemical families and a wide temperature range, from melting point to critical temperature. We demonstrated the efficiency of the introduced models in estimating vaporization enthalpy as well as solvation free energy, as the other important and extensively required thermodynamic quantity associated with vapor-liquid phase change. For describing temperature dependence of vaporization enthalpy, we developed theoretically derived correlations that employed the approximations of the thermodynamically effective molecular surfaces recently conceptualized by us. In addition to the correlations, we also developed highly efficient machine-learning models which yielded the most accurate ever reported results for evaluation of vaporization enthalpy and solvation free energy. We provided a C++ code for a user-friendly and convenient application of the developed machine-learning models.

7.1 Introduction

Evaluation of thermodynamic quantities in the liquid phase is extensively required in numerous scientific and technological applications, ranging from drug discovery [1-3], elucidating the pathway of diseases [4,5], battery design [6,7], nanotechnology [8,9], to the scientific attempts aiming at suppressing the COVID-19 pandemic [10-13] as one of the main global concerns in the present year. Due to the complexity of direct evaluation of thermodynamic quantities in the solution phase, estimations of them through thermodynamic cycles is more commonly encountered in theoretical studies [14-18]. To that end, the most commonly considered thermodynamic quantities are solvation free energy and vaporization enthalpy, due to their convenience of experimental measurement as well as a full characterization of the thermodynamics of the phase change by them. As a complement to our recent study where we investigated the evaluation of solvation free energy [19], in the present study, we aim at precise evaluation of vaporization enthalpy and demonstrating its efficiency in estimating other thermodynamic quantities.

Vaporization enthalpy is one of the key properties of chemicals with numerous widely-required scientific and industrial applications, including but not limited to estimation of saturation vapor pressure (highly required in major chemical processes such as distillation,

evaporation, drying, humidification, and dehumidification [20-22]), calculation of the Hildebrand solubility parameter (required for evaluation of liquid-liquid equilibriums as well as evaluation of the solubility of solids, gases and other liquids [23-25]), evaluation of fire hazards [26], prediction of miscibility of polymer blends as a function of temperature and calculations of liquid-liquid separation processes such as leaching [27,28].

Experimental measurement of vaporization enthalpy for a wide temperature range is not always feasible, e.g., due to safety concerns or operational limitations. This has been the motivation of numerous scientific works in the past decades, aiming at predicting vaporization enthalpy at various temperatures.

Although various approaches such as computer simulation via molecular dynamics [29] or Monte Carlo [30] algorithms or group contributions and QSPR based models [31] have been employed in a number of studies for this purpose, predictive correlations have gained the highest popularity due to offering the most straightforward methods for reliably predicting vaporization enthalpy. The most successful correlations typically predict vaporization enthalpy as a function of temperature and additional more readily available thermophysical properties.

One of the earliest successful models for describing temperature dependence of vaporization enthalpy was developed by Carruth and Kobayashi [32]. They proposed the following relationship for predicting vaporization enthalpy via critical pressure (T_c) and acentric factor (ω):

$$\frac{\Delta H_{vap}}{R\,T_c} = 7.08 \times \left(1 - \frac{T}{T_c}\right)^{0.354} + 10.95 \times \omega \times \left(1 - \frac{T}{T_c}\right)^{0.456}. \tag{7.1}$$

Fish and Lielmezs [33] developed the following model which predicts vaporization enthalpy via normal boiling and critical temperatures as well as the vaporization enthalpy at the normal boiling point (ΔH_{nbp}):

$$\frac{\Delta H_{vap}}{\Delta H_{nbp}} = \frac{T}{T_{nbp}} \times \frac{X + X^q}{1 + X^p}, \tag{7.2}$$

where

$$X = \frac{T}{T_{nbp}} \times \frac{1 - \frac{\mathrm{T}}{T_c}}{1 - \frac{T_{nbp}}{T_c}}, \tag{7.3}$$

and q and p are global parameters, with values of 0.35298 and 0.13856, respectively (except for liquid metals and quantum liquids).

In 1984, Sivaraman et al. [34] proposed their two-reference model written as:

$$\frac{\Delta H_{vap}}{R\,T_c} = \left(\frac{\Delta H_{vap}}{R\,T_c}\right)^{(R1)} + \left(\frac{\omega - 0.21}{0.25}\right) \times \left[\left(\frac{\Delta H_{vap}}{R\,T_c}\right)^{(R2)} - \left(\frac{\Delta H_{vap}}{R\,T_c}\right)^{(R1)}\right], \tag{7.4}$$

where

$$\begin{aligned}
\left(\frac{\Delta H_{vap}}{R\,T_c}\right)^{(R1)} &= 6.537 \times \left(1 - \frac{T}{T_c}\right)^{\frac{1}{3}} - 2.467 \times \left(1 - \frac{T}{T_c}\right)^{\frac{5}{6}} - 77.251 \times \left(1 - \frac{T}{T_c}\right)^{1.208} \\
&\quad + 59.634 \times \left(1 - \frac{T}{T_c}\right) + 36.009 \times \left(1 - \frac{T}{T_c}\right)^{2} - 14.606 \times \left(1 - \frac{T}{T_c}\right)^{3}, \\
\left(\frac{\Delta H_{vap}}{R\,T_c}\right)^{(R2)} - \left(\frac{\Delta H_{vap}}{R\,T_c}\right)^{(R1)} &= -0.133 \times \left(1 - \frac{T}{T_c}\right)^{\frac{1}{3}} - 28.215 \times \left(1 - \frac{T}{T_c}\right)^{\frac{5}{6}} - 82.958 \times \left(1 - \frac{T}{T_c}\right)^{1.208} \\
&\quad + 99.000 \times \left(1 - \frac{T}{T_c}\right) + 19.105 \times \left(1 - \frac{T}{T_c}\right)^{2} - 2.796 \times \left(1 - \frac{T}{T_c}\right)^{3},
\end{aligned} \tag{7.5}$$

Morgan and Kobayashi [35] proposed the following correlation:

$$\Delta H_{vap} = \Delta H_{vap}{}^{(0)} + \omega \Delta H_{vap}{}^{(1)} + \omega^2 \Delta H_{vap}{}^{(2)}, \tag{7.6}$$

where $\Delta H_{vap}{}^{(j)}$ follows the following form proposed by Torquato and Stell [36]:

$$\begin{aligned}
\frac{\Delta H_{vap}{}^{(j)}}{R\,T_c} &= b_1^{(j)}\left(1 - \frac{T}{T_c}\right)^{0.3333} + b_2^{(j)}\left(1 - \frac{T}{T_c}\right)^{0.8333} + b_3^{(j)}\left(1 - \frac{T}{T_c}\right)^{1.2083} \\
&\quad + b_4^{(j)}\left(1 - \frac{T}{T_c}\right) + b_5^{(j)}\left(1 - \frac{T}{T_c}\right)^{2} + b_6^{(j)}\left(1 - \frac{T}{T_c}\right)^{3},
\end{aligned} \tag{7.7}$$

and the eighteen constants $b_i^{(j)}$ are global parameters and their values are reported in table 7.1.

Morgan [56] proposed the following correlation to provide an estimation of vaporization enthalpy via acentric factor:

Table 7.1 Parameters of Morgan and Kobayashi correlation.

	$b_1^{(j)}$	$b_2^{(j)}$	$b_3^{(j)}$	$b_4^{(j)}$	$b_5^{(j)}$	$b_6^{(j)}$
$j = 1$	5.2804	12.865	1.171	-13.116	0.4858	-1.088
$j = 2$	0.80022	273.23	465.08	-638.51	-145.12	74.049
$j = 3$	7.2543	-346.45	-610.48	839.89	160.05	-50.711

$$\frac{\Delta H_{vap}}{R\,T_c} = d_1 \times \left(1 - \frac{T}{T_c}\right)^{d_2 + d_3 \times \frac{T}{T_c} + d_4 \times \left(\frac{T}{T_c}\right)^2}, \tag{7.8}$$

in which

$$\begin{aligned} d_1 &= 7.8149 + 11.409\,\omega + 2.1674\,\omega^2 - 0.65342\,\omega^3, \\ d_2 &= 0.81892 - 0.67637\,\omega + 1.2798\,\omega^2 - 0.47594\,\omega^3, \\ d_3 &= -0.84408 + 1.8297\,\omega - 3.2435\,\omega^2 + 1.1449\,\omega^3, \\ d_4 &= 0.41923 - 1.0892\,\omega + 1.9138\,\omega^2 - 0.65758\,\omega^3. \end{aligned}$$

Our evaluation of the performance of the models mentioned above can be seen in table 7.3 in the results section below.

Despite several decades of research on this specific topic, almost all of the successful correlations applicable for predicting vaporization enthalpy of diverse compounds at wide temperature ranges have mainly been limited to empirically developed models. That motivated us to develop and benchmark a correlation entirely on a theoretical basis in our recent studies, which describes the temperature dependence of vaporization enthalpy as [38,39]:

$$\Delta H_{vap} = \frac{a_s \gamma^\circ}{2}\left(2\left(1 - \frac{T}{T_c}\right)^{11/9} + \frac{11}{9}\frac{T}{T_c}\left(1 - \frac{T}{T_c}\right)^{2/9}\right) - \frac{R}{2}\,T.\ln\left(\frac{T}{T_c}\right). \tag{7.9}$$

In Eq. (7.9), the parameter α_s is our recently conceptualized *thermodynamically effective* surface of molecules [38] and $\gamma^\circ$is the pre-factor of the Guggenheim–Katayama relationship [40] employed for describing the temperature dependence of surface tension. As we demonstrated in our recent study [38], with an appropriate estimation of α_s, a reliable and

highly accurate evaluation of not only vaporization enthalpy but also other thermodynamics quantities is achievable.

In the present study, we investigate practical and straightforward approximations of the thermodynamically effective surfaces and show advantages of predicting vaporization enthalpy through the resulting relationships, compared to other empirically proposed correlations.

Generally, approximations of thermodynamically effective surfaces, as required by Eq. (7.9), can be achieved by three different methods. The first method is approximating those surfaces using experimental data of molar volume at a specific temperature and assuming molecules as perfect spheres [38]. This method has been one of the earliest examples of approximating molecular surfaces via experimental data and was initially suggested by Eötvös in 1886 [41]. The second method for approximating molecular surfaces is by exploiting the well-established computer algorithms developed to calculate vdW or solvent excluded surfaces of molecules [38]. Nevertheless, diversity in approximated molecular surfaces obtained via different parameterizations of atomic radii in those methods can yield quite different values for the predicted vaporization enthalpy and hence requires careful attention, as we demonstrated in our recent study [38]. Finally, the thermodynamically effective molecular surfaces can be approximated using reference data of vaporization enthalpy at a single or multiple arbitrary temperatures. Similarly, the pre-factor $\frac{a_s\gamma^\circ}{2}$ in Eq. (7.9) can be considered as an adjustable parameter and its determination can then be achieved via single or multiple reference data of vaporization enthalpy.

In the present study, we employ approximations of molecular surfaces based on the third method to develop practical and straightforward correlations, while the molecular surfaces approximated via the first two methods will be employed for developing machine-learning models. Accordingly, we employ machine learning for mapping the dependency between vaporization enthalpy and a number of potentially relevant quantities, most specifically appropriate approximations of thermodynamically effective surfaces obtained via efficient computer algorithms or molar volume as discussed above. As we have shown in our recent study, the thermodynamically effective surfaces are generally very similar to the van-der-Waals (vdW) surfaces of molecules [38]. Nevertheless, even the slight deviation which exists between the two surfaces can have significant impacts on the accuracy of evaluated thermodynamic quantities. Therefore, the main role of machine learning here is mostly to learn and modify the impacts of such inaccuracies on estimating molecular surfaces.

Noteworthy, employing machine learning to predict vaporization enthalpy via other more readily available quantities has also been studied elsewhere. Nevertheless, the available machine-learning models proposed in the literature are mainly limited to single-point vapor-

ization enthalpy prediction for either room temperature [42,43] or the normal boiling point [44,45]. In contrast, the machine-learning models proposed in the present study are capable to evaluate vaporization enthalpy at any temperature of interest, from melting point to the critical temperature.

7.2 Computational details

7.2.1 Dataset

The performance of the models studied in the current work was benchmarked using the thermophysical data of the DIPPR801 database [46]. Screening the dataset and selecting only the compounds with a maximum uncertainty of 5% results in 828 compounds from diverse chemical families. The names of these compounds are provided as supplementary material.

For each compound, the experimentally determined data of vaporization enthalpies for 25 points linearly distributed between the melting point and the critical temperature approximated via the provided relationships in the DIPPR database were employed.

The performance of vaporization enthalpy predictive correlations is reported as Average Absolute Deviation (AAD) defined as:

$$\mathrm{AAD} = \frac{1}{N} \sum \left(\left| y_i^{exp} - y_i^{pred} \right| \right), \tag{7.10}$$

as a more appropriate parameter to evaluate the performance of the models compared to relative error. This is because at temperatures close to the critical point, the vaporization enthalpy approaches zero and as a result, small deviations in predicted data yield very large relative errors, leading to inappropriate inferences about the performance of the studied models.

7.2.2 Machine learning

In addition to the proposed theoretically derived correlations, we also studied the predictability of the vaporization enthalpy via machine learning. To that end, we employed artificial neural networks to map the dependency between the vaporization enthalpy and a number of input variables. We studied 9 potentially relevant thermophysical properties, namely critical temperature (T_c), critical pressure (P_c), normal boiling point (NBP), liquid molar volume (V), acentric factor (ω), radius of gyration (R_g), van-der-Waals area (vdW_A) and volume (vdW_V) and dielectric constant (ε). Among the considered quantities, the first five are commonly

employed in vaporization enthalpy predictive models, as discussed in the introduction section. Among the other quantities, vdW parameters and radius of gyration were selected due to being directly related to molecular surfaces, and the dielectric constant was considered due to having a direct impact on solution thermodynamics [47].

After trying various models, we noticed that the most successful neural network models are achieved for non-dimensionalized target and input variables. Accordingly, we employed reduced temperature ($\frac{T}{T_c}$) as an additional model input and $\frac{\Delta H_{vap}}{RT_c}$ as the target variable.

Considering that for many compounds not all of the model inputs might be readily available, we only studied models which required 2 to 6 of the potentially related quantities stated above. To that end, we exploited the Minimum Redundancy and Maximum Relevance (MRMR) algorithm [48] as a highly efficient algorithm for selecting the most effective sets of variables for developing robust machine-learning-based models [49]. This produced 23 different combinations of input variables, listed in table 7.2. The screened variables in addition to computed molecular surfaces, discussed in the next section, were considered as the inputs of the neural network models. The schematic representation of the studied neural network is depicted in figure 7.1.

Developing neural network models was carried out based on the guidelines proposed in our previous studies [47,50]. Accordingly, we used only 60% of the dataset for training the neural network models and assigned 15% and 25% of the dataset for validation and test of the models, respectively. We only studied neural networks with one hidden layer. For each model, the upper limit for the numbers of neurons was set in a way to allow existence of roughly 10 training samples or more per model constant, as recommended in the previous study [50]. For each model, we considered Levenberg-Marquardt backpropagation and gradient descent backpropagation training algorithms, and hidden layer transfer functions of the logarithm-sigmoid and tangent-sigmoid types [51]. Although initial training of the models was carried out by cross-validation mentioned above, for models which yielded the highest accuracy their reliability was further validated by a rigorous 100-fold cross-validation. Accordingly, those selected models were retrained for 100 different randomly divided training, validation, and test sets employing the previously optimized weight and bias constants as the initial guesses. The models which in at least 80% of the iterations yielded test and training set errors with the same mean and standard deviation as evaluated by the t-test method were selected as reliably trained models.

7.2.3 Computation of molecular surfaces

In addition to the thermophysical quantities discussed in the previous section, the molecular surfaces evaluated via molar volumes as well as vdW surfaces were considered as additional

Table 7.2 Various combinations of screened thermophysical quantities in selected machine-learning models.

Model ID	**Model inputs:**					
1	T/T_c	T_c	ω	vdW_A	ε	$\Delta H_{nbp}/RT_c$
2	T/T_c	ω	vdW_A	ε	$\Delta H_{nbp}/RT_c$	
3	T/T_c	NBP	V	ω	ε	$\Delta H_{nbp}/RT_c$
4	T/T_c	NBP	ω	ε	$\Delta H_{nbp}/RT_c$	
5	T/T_c	ω	ε	$\Delta H_{nbp}/RT_c$		
6	T/T_c	NBP	ω	$\Delta H_{nbp}/RT_c$		
7	T/T_c	P_c	NBP	ε	$\Delta H_{nbp}/RT_c$	
8	T/T_c	ω	$\Delta H_{nbp}/RT_c$			
9	T/T_c	T_c	NBP	ω	vdW_A	ε
10	T/T_c	NBP	V	ω	vdW_V	ε
11	T/T_c	T_c	NBP	ω	ε	
12	T/T_c	ω	vdW_A	ε		
13	T/T_c	Pc	NBP	ω	ε	
14	T/T_c	T_c	P_c	ω	ε	
15	T/T_c	NBP	ω	ε		
16	T/T_c	Pc	ω	ε		
17	T/T_c	ω	ε			
18	T/T_c	Pc	ω			
19	T/T_c	NBP	V	ω		
20	T/T_c	T_c	NBP	ω		
21	T/T_c	NBP	ω			
22	T/T_c	ω				
23	T/T_c	T_c	P_c	NBP	vdW_A	ε

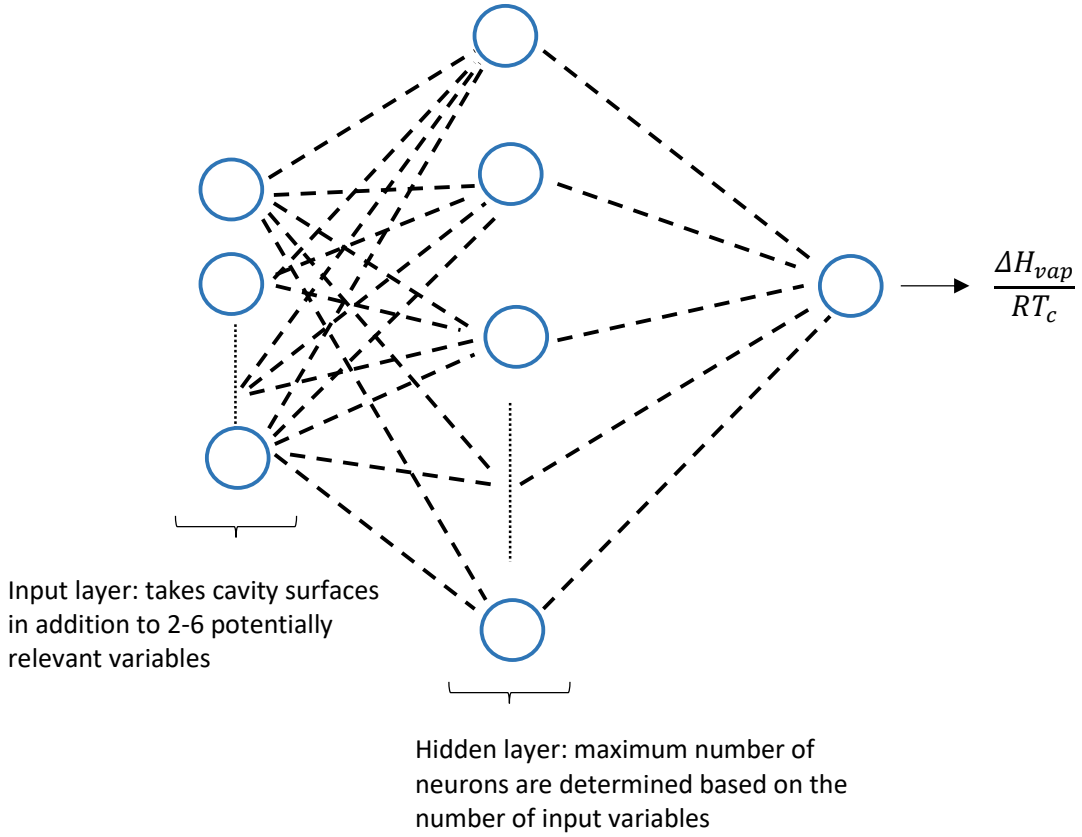


Fig. 7.1 Schematic view of the studied neural network models.

input variables for the studied machine-learning models. For approximating the molecular surfaces via the former method, the experimental data of molar volume at the melting and normal boiling points were considered separately.

The vdW surfaces were computed in Gaussian 16 using the GEPOL algorithm [52]. To that end, the geometry of each molecule was first optimized at the B3LYP/6-311+G(2d,p) level of theory. The optimized geometries were then used for calculation of the vdW surfaces within the SCRF module of Gaussian 16 based on the Bondi parameterization of atomic radii, as an appropriate parameterization of these surfaces [38]. Examples of molecular surfaces calculated based on the above recipe for water, benzene, and ammonia, are depicted in figure 7.2.

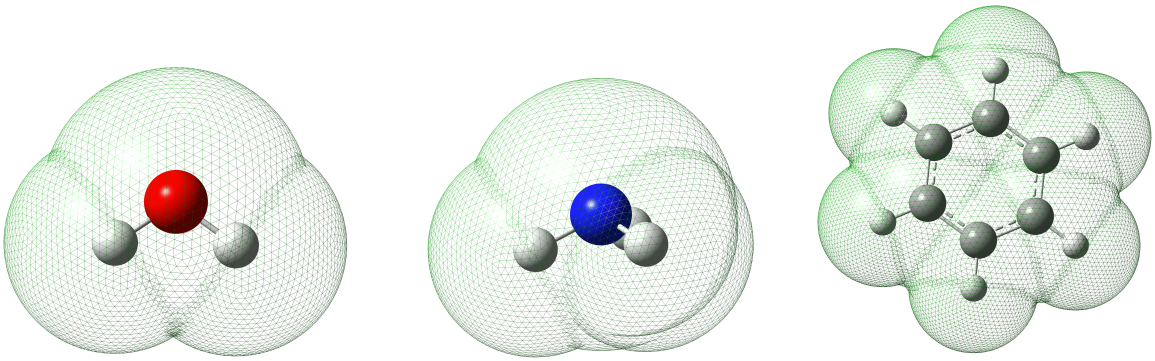


Fig. 7.2 The vdW surfaces of water, ammonia and benzene based on the Bondi parameterization.

7.3 Results and discussion

7.3.1 Evaluation of vaporization enthalpy

For a straightforward and practical evaluation of vaporization enthalpy via Eq. (7.9), we employed approximations of $\frac{a_s\gamma^\circ}{2}$ as an adjustable parameter using a single reference data point of vaporization enthalpy. As discussed earlier, prediction of vaporization enthalpy using Eq. (7.9) can also be achieved by approximating a_s via molar volumes and γ° via surface tension data [38]. Nevertheless, approximating $\frac{a_s\gamma^\circ}{2}$ using a single vaporization enthalpy data point possesses the advantage of allowing its application for compounds for which the surface tension data and the γ° constant are not available.

We studied approximations of the $\frac{a_s\gamma^\circ}{2}$ parameter determined via a single reference data point of vaporization enthalpy at the normal boiling point and also at the melting point.

For the $\frac{a_s\gamma^\circ}{2}$ parameter approximated using experimentally determined vaporization enthalpy at normal boiling point and melting point, we obtained AAD of 1185.5, and 1406.3 J/mol, respectively. A comparison of these results with those obtained through the most successful empirically developed correlations, reported in table 7.3, shows a comparable or better accuracy for our proposed approach. Nevertheless, it should be noted that due to the uncertainty in the accuracy of reference data which for our studied dataset can be up to 5 percent as reported by DIPPR, slight differences in obtained accuracies for different models do not allow a certain judgment about their performances. Specifically, most of the empirically developed models are parameterized to reproduce DIPPR data, which might result in higher accuracies for those models. For example, the higher accuracy of the model developed by Morgan [56] compared to other empirical models might be due to employing an algebraic relationship in that model exactly similar to the one employed by the DIPPR database to provide reference data at various temperatures.

In addition to the experimentally determined data of vaporization enthalpy at normal boiling and melting points, we also studied approximations of the $\frac{a_s\gamma^\circ}{2}$ parameter using estimations of vaporization enthalpy at normal boiling point via three reliable and well-benchmarked models proposed for this purpose by Chen [53], Vetere [54] and Liu [55]. These models provide accurate predictions of ΔH_{nbp} using critical temperature and critical pressure and normal boiling point. Using these approximations of the $\frac{a_s\gamma^\circ}{2}$ parameter, we obtained AADs of 1408, 1586.9, and 1497.4 J/mol, respectively, for predicted vaporization enthalpies of all other temperatures.

The most eminent advantage of employing this latter approach for predicting vaporization enthalpy is that it allows straightforward, highly practical, and yet accurate evaluation of vaporization enthalpy only via critical temperature, critical pressure, and normal boiling, which is not possible by any other model proposed elsewhere, to the best of our knowledge.

Note that, as discussed before and as it can be seen in table 7.3, the most accurate empirically developed models typically require the acentric factor for predicting vaporization enthalpy. Nevertheless, experimental measurement of the acentric factor itself requires accurate measurement of vapor pressure at 0.7 T_c as well as accessing both critical pressure and temperature which can be more challenging than measuring vaporization enthalpy itself.

In addition to the correlations, we also investigated the predictability of vaporization enthalpy via employing approximated molecular surfaces through machine learning. The results obtained via the studied machine-learning models are reported in table 7.4.

According to the results, based on the number and type of input variables, the developed machine-learning models yield different AADs ranging from 487.717 to 1822.17 J/mol.

As implied from these results, our developed machine-learning models provide the highest accuracy ever reported for the prediction of vaporization enthalpy as well as significant flexibility in choosing diverse sets of input variables. Additionally, for the same input variables, the results obtained via the machine-learning models still show much higher accuracy compared to the most successful predictive correlations, which are the outcome of several decades of efforts in developing such correlations. For example, for predicting the vaporization enthalpy via critical temperature and acentric factor as the only model inputs, the neural network model specified with ID 22 in table 7.4 can be used, which results in an AAD of 992.158 J/mol without requiring any additional parameter. If we consider molecular surfaces as an additional input variable, then we can further improve the accuracy to 682.054 J/mol. This result shows a significantly higher accuracy compared to the predictive correlations reported in table 7.3 which require the same model inputs. Among the developed machine-learning models, one of the most interesting results is those obtained via a model specified in table 7.3 with ID 23, which requires only experimental data at the critical point

Table 7.3 Comparison of the results predicted via various models.

	Model inputs	AAD (J/mol)
Machine learning	See table 7.4 for details	487.717-1822.17
New relationship (eq. (7.9)) (α_s determined via $\Delta H_{nbp,exp}$)	$\Delta H_{nbp}, T_c$	1185.5
New relationship (eq. (7.9)) (α_s determined via $\Delta H_{melp,exp}$)	$\Delta H_{melt}, T_c$	1406.3
New relationship (eq. (7.9)) (α_s determined via Chen model [53])	T_{nbp}, T_c, P_c	1408.1
New relationship (eq. (7.9)) (α_s determined via Vetere model [54])	T_{nbp}, T_c, P_c	1497.4
New relationship (eq. (7.9)) (α_s determined via Liu model [55])	T_{nbp}, T_c, P_c	1586.9
Fish-Lielmezs [33]	$T_{nbp}, T_c, \Delta H_{nbp}$	1365.4
Morgan [56]	T_c, ω	1108.9
Morgan-Kobayashi [35]	T_c, ω	1312.9
Sivaraman et al. [34]	T_c, ω	1314.7
Carruth-Kobayashi [32]	T_c, ω	1444.4
Meyra et al. [37]	$T_{nbp}, T_c, \Delta H_{nbp}$	2122.1

Table 7.4 AAD (J/mol) obtained through studied developed machine-learning models and different estimations of molecular surfaces.

Model ID	Thermo.	Thermo.- S(vdW)	Thermo.- S(melt)	Thermo.- S(NBP)
1	865.383	487.717	505.281	497.931
2	867.641	508.667	513.841	515.912
3	868.492	484.070		
4	869.496	485.354	520.522	501.634
5	890.629	514.273	537.269	522.163
6	897.208	489.808	520.547	501.845
7	912.854	586.567	602.086	600.743
8	913.738	520.436	554.833	538.750
9	945.200	656.788	662.974	659.674
10	946.728	658.542		
11	952.052	658.063	675.586	658.662
12	952.575	666.334	687.141	685.864
13	957.393	647.452	614.900	610.027
14	958.143	602.940	637.400	604.333
15	961.256	659.526	690.667	675.295
16	961.524	655.665	687.521	670.938
17	970.587	674.591	690.769	689.173
18	977.585	678.750	698.751	695.268
19	980.281	671.093		
20	981.209	658.938	688.263	687.942
21	984.434	680.168	696.705	696.726
22	992.158	682.054	696.690	696.605
23	997.771	610.200	708.442	677.050

Thermo.: the thermophysical quantities introduced in table 7.2
Thermo.- S(vdW): thermophysical quantities and vdW surfaces are used as model input
Thermo.- S(melt): thermophysical quantities and surfaces evaluated via molar volume at melting point are used as model input
Thermo.- S(NBP): thermophysical quantities and surfaces evaluated via molar volumes at NBP are used as model input

and the dielectric constant and yields an AAD as low as 610.2 J/mol. In comparison to all predictive correlations reported in table 7.3, this model shows an obvious advantage because all those correlations in addition to experimentally determined data at the critical point also require at least one additional reference data item of vaporization enthalpy either at 0.7 T_c (required for calculating the acentric factor) or at normal boiling point, nevertheless resulting in lower accuracy compared to the machine-learning model.
A schematic comparison of the performance of different vaporization enthalpy predictive models for the 6 most widely employed solvents is provided in figure 7.3.

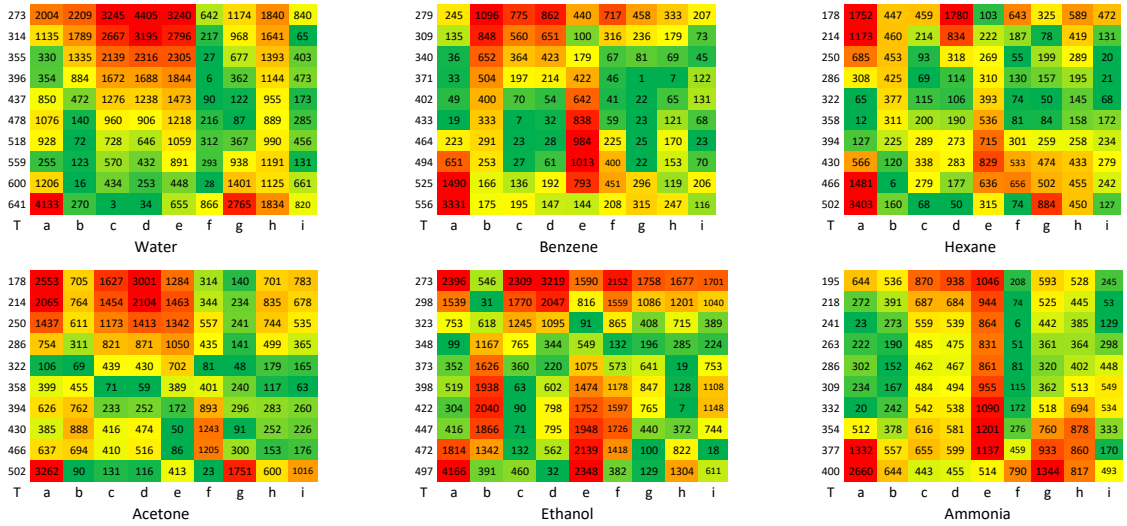


Fig. 7.3 Schematic comparison of AAD obtained for different temperatures specified in the first column via the predictive models: (a) Eq. (7.9) and molecular surfaces evaluated via ΔH_{nbp}, (b) Carruth-Kobayashi, (c) Morgan, (d) Morgan- Kobayashi , (e) Sivaraman, (f) Fish_Lielmezs, (g) ML-1 , (h) ML-22, (i) ML-23.

As can be seen in figure 7.3, the machine-learning models show in general the most reliable evaluation of vaporization enthalpy for the whole temperature range.

7.3.2 Estimation of other thermodynamic quantities through the proposed models

One of the main advantages of studying the temperature dependence of vaporization enthalpy is the provided possibility to estimate all other thermodynamic quantities encountered in vapor-liquid phase change through them. As the most important example, here we demonstrate evaluation of solvation free energy, as one of the most extensively applied thermodynamic quantities. Accurate evaluation of solvation free energy is highly required

by a very broad range of scientific fields, including but not limited to life sciences [57-60], nanotechnology [61,62], energy storage [63-65], and many other cutting-edge applications. Estimation of solvation free energy through temperature dependence of vaporization enthalpy can be achieved via the Gibbs-Helmholtz relationship defined as:

$$\Delta G_{solvation} = -T \int \frac{\Delta H_{vap}}{T^2}\, dT\,, \tag{7.11}$$

where the constant of integration can be obtained via [38,66]:

$$P_{atm} = \frac{R\, T_{nbp}}{V_{m,nbp}} \exp(\frac{\Delta G_{sol,nbo}}{K T_{nbp}})\,. \tag{7.12}$$

Here, $\Delta G_{sol,nbo}$ which is the solvation free energy at the normal boiling point, is calculated via atmospheric pressure (P_{atm}), liquid molar volume at the normal boiling point ($V_{m,nbp}$) and normal boiling point temperature (T_{nbp}). Via the abovementioned calculations, we studied the predictability of solvation free energies for the experimental data provided by the Minnesota solvation database [67]. To that end, the vaporization enthalpies predicted via machine-learning models with IDs 1, 22, and 23 were converted to fourth-order polynomials by curve fitting and were used to analytically solve the integral in Eq. (7.11).

Via this recipe, we estimated the solvation free energies for 30 compounds which were common between our studied database and the Minnesota solvation database. The obtained results are compared with the experimental data in table 7.5. According to these results, via the vaporization enthalpies predicted by the proposed machine-learning models with IDs 1, 22, and 23, we could reproduce experimentally determined data of solvation free energies with AADs of 0.087667, 0.10267, 0.095333 kcal/mol, respectively. According to our recent review of the state-of-the-art theoretical and empirical approaches applicable to the evaluation of solvation free energy [19], by at least a factor of two, the newly proposed machine-learning models provide the most accurate evaluation of solvation free energy.

7.4 Conclusion

In the present study, we introduced practical approaches for an accurate evaluation of thermodynamic quantities associated with vapor-liquid phase change. We demonstrated the success of our introduced models in the high accuracy estimation of vaporization enthalpy and solvation free energy. In addition to correlations that provided a straightforward evaluation of

Table 7.5 Comparison of experimentally determined and evaluated solvation free energies obtained through temperature dependence of vaporization enthalpy and proposed machine learning models.

Compound	Exp.	ML-1	ML-22	ML-23
2-MethylPyridine	-5.71	-5.79	-5.79	-5.81
Acetonitrile	-4.85	-4.91	-4.99	-4.92
AcetoPhenone	-7.59	-7.79	-7.77	-7.80
Aniline	-7.61	-7.58	-7.52	-7.60
Anisole	-6.33	-6.45	-6.43	-6.45
Benzene	-4.55	-4.57	-4.57	-4.57
BromoEthane	-3.67	-3.70	-3.71	-3.70
Bromoform	-6.21	-6.33	-6.30	-6.31
ChloroBenzene	-5.66	-5.77	-5.77	-5.79
Chloroform	-4.13	-4.19	-4.20	-4.19
CycloHexane	-4.43	-4.43	-4.44	-4.43
CycloHexanone	-6.25	-6.38	-6.39	-6.40
2,6-DiMethylPyridine	-6.04	-6.13	-6.08	-6.11
Ethanol	-5.04	-5.10	-5.11	-5.10
EthylBenzene	-5.67	-5.76	-5.77	-5.78
FluoroBenzene	-4.60	-4.66	-4.65	-4.66
2-Propanol	-4.82	-5.10	-5.02	-5.03
IsoPropylBenzene	-6.04	-6.11	-6.11	-6.12
m-Cresol	-8.40	-8.29	-8.09	-8.15
Mesitylene	-6.40	-6.53	-6.55	-6.56
Dichloromethane	-3.80	-3.84	-3.83	-3.83
NitroEthane	-5.53	-5.62	-5.71	-5.66
NitroMethane	-5.38	-5.47	-5.55	-5.48
Pyridine	-5.47	-5.55	-5.53	-5.56
tert-ButylBenzene	-6.43	-6.44	-6.39	-6.40
TetraHydroFuran	-4.25	-4.29	-4.29	-4.29
Toluene	-5.12	-5.19	-5.20	-5.21
TriEthylAmine	-4.44	-4.49	-4.49	-4.48
1,2,4-TriMethylBenzene	-6.47	-6.65	-6.67	-6.67
Water	-6.31	-6.44	-6.48	-6.42

thermodynamic quantities, we also developed machine-learning models which, unlike other machine-learning models proposed elsewhere, are not limited to a single temperature and can be used to predict vaporization enthalpy at various temperatures. We emphasize significant advantages of machine-learning models: They possess higher flexibility in input variables, which for some models are more readily available, and yield the highest ever reported accuracy for prediction of vaporization enthalpy as a function of temperature and solvation free energy. For the demonstrated advantage of machine-learning models compared to empirical correlations in the evaluation of vaporization enthalpy, one limitation in employing those models is the requirement of having technical knowledge for their application. With that in mind, in the supplementary material, we provide a C++ code with detailed instructions for a user-friendly and straightforward application of the developed machine-learning models.

References:

1. Kundu, D.; Dubey, V. K., Potential alternatives to current cholinesterase inhibitors: An in silico drug repurposing approach. Drug Development and Industrial Pharmacy 2021, (just-accepted), 1-30.
2. Adeniyi, A. A.; Conradie, J., Computational insight into the anticholinesterase activities and electronic properties of physostigmine analogs. Future medicinal chemistry 2019, 11 (16), 1907-1928.
3. Magistrato, A.; Sgrignani, J.; Krause, R.; Cavalli, A., Single or multiple access channels to the CYP450s active site? An answer from free energy simulations of the human aromatase enzyme. The journal of physical chemistry letters 2017, 8 (9), 2036-2042.
4. Serrano-Aparicio, N.; Moliner, V.; Swiderek, K., Nature of Irreversible Inhibition of Human 20S Proteasome by Salinosporamide A. The Critical Role of Lys–Asp Dyad Revealed from Electrostatic Effects Analysis. ACS Catalysis 2021, 11 (6), 3575-3589.
5. Leonard, C.; Phillips, C.; McCarty, J., Insight Into Seeded Tau Fibril Growth From Molecular Dynamics Simulation of the Alzheimer's Disease Protofibril Core. Frontiers in molecular biosciences 2021, 8, 109.
6. Choi, Y.; Preston, T. J.; Adamczyk, A. J., Data-driven investigation of monosilane and ammonia copyrolysis to silicon-nitride-based ceramic nanomaterials. ChemPhysChem 2020, 21 (22), 2627-2642.
7. Lener, G.; Otero, M.; Barraco, D.; Leiva, E. P. M., Energetics of silica lithiation and its applications to lithium ion batteries. Electrochimica Acta 2018, 259, 1053-1058.
8. Tong, W.-Y.; Zhao, T.-T.; Zhao, X.-F.; Wang, X.; Wu, Y.-B.; Yuan, C., Neutral nano-polygons with ultrashort Be–Be distances. Dalton Transactions 2019, 48 (42), 15802-15809.
9. El-Baba, T. J.; Clemmer, D. E., Solution thermochemistry of concanavalin A tetramer conformers measured by variable-temperature ESI-IMS-MS. International journal of mass spectrometry 2019, 443, 93-100.
10. Pang, J.; Gao, S.; Sun, Z.; Yang, G., Discovery of small molecule PLpro inhibitor against COVID-19 using structure-based virtual screening, molecular dynamics simulation, and molecular mechanics/Generalized Born surface area (MM/GBSA) calculation. Structural chemistry 2021, 32 (2), 879-886.
11. Ding, H.-m.; Yin, Y.-w.; Sheng, Y.-j.; Ma, Y.-q., Accurate Evaluation on the Interactions of SARS-CoV-2 with Its Receptor ACE2 and Antibodies CR3022/CB6. Chinese Physics Letters 2021, 38 (1), 018701.
12. Liu, J.; Zhai, Y.; Liang, L.; Zhu, D.; Zhao, Q.; Qiu, Y., Molecular modeling evaluation of the binding effect of five protease inhibitors to COVID-19 main protease. Chemical Physics 2021, 542, 111080.
13. Li, Z.; Li, X.; Huang, Y.-Y.; Wu, Y.; Liu, R.; Zhou, L.; Lin, Y.; Wu, D.; Zhang, L.; Liu, H., Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual

screening of existing drugs. Proceedings of the National Academy of Sciences 2020, 117 (44), 27381-27387.
14. Fowles, D. J.; Palmer, D. S.; Guo, R.; Price, S. L.; Mitchell, J. B., Toward Physics-Based Solubility Computation for Pharmaceuticals to Rival Informatics. Journal of chemical theory and computation 2021.
15. Itkis, D.; Cavallo, L.; Yashina, L. V.; Minenkov, Y., Ambiguities in solvation free energies from cluster-continuum quasichemical theory: lithium cation in protic and aprotic solvents. Physical Chemistry Chemical Physics 2021, 23 (30), 16077-16088.
16. Chen, R.; Deng, S.; Xu, W.; Zhao, L., A graphic analysis method of electrochemical systems for low-grade heat harvesting from a perspective of thermodynamic cycles. Energy 2020, 191, 116547.
17. Abraham, N. S.; Shirts, M. R., Statistical mechanical approximations to more efficiently determine polymorph free energy differences for small organic molecules. Journal of Chemical Theory and Computation 2020, 16 (10), 6503-6512.
18. Gapsys, V.; Seeliger, D.; de Groot, B. L., New soft-core potential function for molecular dynamics based alchemical free energy calculations. Journal of chemical theory and computation 2012, 8 (7), 2373-2382.
19. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. 2021.
20. Verevkin, S. P.; Andreeva, I. V.; Pimerzin, A. A., Evaluation of vaporization thermodynamics of pure amino-alcohols. Journal of Molecular Liquids 2021, 335, 116568.
21. Zaitsau, D. H.; Verevkin, S. P., Vaporization Thermodynamics of Morpholinium Based Ionic Liquids. Zeitschrift für anorganische und allgemeine Chemie 2021, 647 (5), 547-551.
22. Zaitsau, D. H.; Plechkova, N.; Verevkin, S. P., Vaporization thermodynamics of ionic liquids with tetraalkylphosphonium cations. The Journal of Chemical Thermodynamics 2019, 130, 204-212.
23. Albahri, T. A., Accurate prediction of the solubility parameter of pure compounds from their molecular structures. Fluid Phase Equilibria 2014, 379, 96-103.
24. Weerachanchai, P.; Chen, Z.; Leong, S. S. J.; Chang, M. W.; Lee, J.-M., Hildebrand solubility parameters of ionic liquids: Effects of ionic liquid type, temperature and DMA fraction in ionic liquid. Chemical engineering journal 2012, 213, 356-362.
25. Rai, N.; Wagner, A. J.; Ross, R. B.; Siepmann, J. I., Application of the TraPPE force field for predicting the Hildebrand solubility parameters of organic solvents and monomer units. Journal of Chemical Theory and Computation 2008, 4 (1), 136-144.
26. Alibakhshi, A.; Mirshahvalad, H.; Alibakhshi, S., A modified group contribution method for accurate prediction of flash points of pure organic compounds. Industrial & Engineering Chemistry Research 2015, 54 (44), 11230-11235.
27. Genheden, S.; Essex, J. W., A simple and transferable all-atom/coarse-grained hybrid model to study membrane processes. Journal of chemical theory and computation 2015, 11 (10), 4749-4759.
28. Yang, L.; Adam, C.; Cockroft, S. L., Quantifying solvophobic effects in nonpolar cohesive interactions. Journal of the American Chemical Society 2015, 137 (32), 10084-10087.
29. Lousada, C. M.; Pinto, S. S.; Canongia Lopes, J. N.; Minas da Piedade, M. F.; Diogo, H. P.; Minas da Piedade, M. E., Experimental and molecular dynamics simulation study of the sublimation and vaporization energetics of iron metalocenes. Crystal structures of (η5-C5H4CH3) 2 and Fe [(η5-(C5H5)(η5-C5H4CHO)]. The Journal of Physical Chemistry A 2008, 112 (13), 2977-2987.
30. Martin, M. G.; Biddy, M. J., Monte Carlo molecular simulation predictions for the heat of vapor-

ization of acetone and butyramide. Fluid phase equilibria 2005, 236 (1-2), 53-57.
31. Abdi, S.; Movagharnejad, K.; Ghasemitabar, H., Estimation of the enthalpy of vaporization at normal boiling temperature of organic compounds by a new group contribution method. Fluid Phase Equilibria 2018, 473, 166-174.
32. Carruth, G. F.; Kobayashi, R., Extension to low reduced temperatures of three-parameter corresponding states: vapor pressures, enthalpies and entropies of vaporization, and liquid fugacity coefficients. Industrial & Engineering Chemistry Fundamentals 1972, 11 (4), 509-517.
33. Fish, L. W.; Lielmezs, J., General method for predicting the latent heat of vaporization. Industrial & Engineering Chemistry Fundamentals 1975, 14 (3), 248-256.
34. Sivaraman, A.; Magee, J. W.; Kobayashi, R., Generalized correlation of latent heats of vaporization of coal-liquid model compounds between their freezing points and critical points. Industrial & engineering chemistry fundamentals 1984, 23 (1), 97-100.
35. Morgan, D. L.; Kobayashi, R., Extension of Pitzer CSP models for vapor pressures and heats of vaporization to long-chain hydrocarbons. Fluid Phase Equilibria 1994, 94, 51-87.
36. Torquato, S.; Stell, G. R., An equation for the latent heat of vaporization. Industrial & Engineering Chemistry Fundamentals 1982, 21 (3), 202-205.
37. Meyra, A. G.; Kuz, V. A.; Zarragoicoechea, G. J., Universal behavior of the enthalpy of vaporization: an empirical equation. Fluid Phase Equilibria 2004, 218 (2), 205-207.
38. Alibakhshi, A., Thermodynamically effective molecular surfaces for more efficient study of condensed-phase thermodynamics. 2021.
39. Alibakhshi, A., Enthalpy of vaporization, its temperature dependence and correlation with surface tension: a theoretical approach. Fluid Phase Equilibria 2017, 432, 62-69.
40. Adam, N., he physics and chemistry of surfaces (3d ed.): Oxford University Press. London: 1941.
41. Eötvös, R., Ueber den Zusammenhang der Oberflächenspannung der Flüssigkeiten mit ihrem Molecularvolumen. Annalen der Physik 1886, 263 (3), 448-459.
42. Jaquis, B. J.; Li, A.; Monnier, N. D.; Sisk, R. G.; Acree, W. E.; Lang, A. S., Using machine learning to predict enthalpy of solvation. Journal of Solution Chemistry 2019, 48 (4), 564-573.
43. Golmohammadi, H.; Dashtbozorgi, Z., Prediction of solvation enthalpy of gaseous organic compounds in propanol. Russian Journal of Physical Chemistry A 2016, 90 (9), 1806-1812.
44. Abooali, D.; Sobati, M. A., Novel method for prediction of normal boiling point and enthalpy of vaporization at normal boiling point of pure refrigerants: A QSPR approach. International journal of refrigeration 2014, 40, 282-293.
45. Mohammadi, A. H.; Richon, D., New predictive methods for estimating the vaporization enthalpies of hydrocarbons and petroleum fractions. Industrial & engineering chemistry research 2007, 46 (8), 2665-2671.
46. Wilding, W. V.; Rowley, R. L.; Oscarson, J. L., DIPPR® Project 801 evaluated process design data. Fluid phase equilibria 1998, 150, 413-420.
47. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. Nature Communications 2021,12(1),1-7.
48. Peng, H.; Long, F.; Ding, C., Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence 2005, 27 (8), 1226-1238.
49. Brown, G. In A new perspective for information theoretic feature selection, Artificial intelligence and statistics, 2009; pp 49-56.

50. Alibakshi, A., Strategies to develop robust neural network models: Prediction of flash point as a case study. Analytica chimica acta 2018, 1026, 69-76.
51. Demuth, H.; Beale, M., Neural Network Toolbox For Use with Matlab–User'S Guide Verion 3.0. 1993.
52. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., Gaussian 16. Revision A 2016, 3.
53. Chen, N., Generalized Correlation for Latent Heat of Vaporization. Journal of Chemical and Engineering Data 1965, 10 (2), 207-210.
54. Vetere, A., New Generalized Correlations for Enthalpy of Vaporization of Pure Compounds. Laboratori Ricerche Chimica Industriale, San Donato Milanese 1973.
55. LIU, Z.-Y., Estimation of heat of vaporization of pure liquid at its normal boiling temperature. Chemical Engineering Communications 2001, 184 (1), 221-228.
56. Morgan, D. L., Use of transformed correlations to help screen and populate properties within databanks. Fluid phase equilibria 2007, 256 (1-2), 54-61.
57. Leonard, A. N.; Lyman, E., Activation of G-protein-coupled receptors is thermodynamically linked to lipid solvation. Biophysical Journal 2021, 120 (9), 1777-1787.
58. Sumi, T.; Imamura, H., Water-Mediated Interactions Destabilize Proteins. Protein Science 2021.
59. Huang, K.; Luo, S.; Cong, Y.; Zhong, S.; Zhang, J. Z.; Duan, L., An accurate free energy estimator: based on MM/PBSA combined with interaction entropy for protein–ligand binding affinity. Nanoscale 2020, 12 (19), 10737-10750.
60. Rifai, E. A.; Ferrario, V.; Pleiss, J. r.; Geerke, D. P., Combined linear interaction energy and alchemical solvation free-energy approach for protein-binding affinity computation. Journal of chemical theory and computation 2020, 16 (2), 1300-1310.
61. Samani, M. T.; Hashemianzadeh, S. M., The effect of functionalization on solubility and plasmonic features of gold nanoparticles. Journal of Molecular Graphics and Modelling 2020, 101, 107749.
62. Wu, J.; Japip, S.; Chung, T.-S., Infiltrating molecular gatekeepers with coexisting molecular solubility and 3D-intrinsic porosity into a microporous polymer scaffold for gas separation. Journal of Materials Chemistry A 2020, 8 (13), 6196-6209.
63. Hao, J.; Yuan, L.; Ye, C.; Chao, D.; Davey, K.; Guo, Z.; Qiao, S. Z., Boosting Zinc Electrode Reversibility in Aqueous Electrolytes by Using Low-Cost Antisolvents. Angewandte Chemie 2021, 133 (13), 7442-7451.
64. Asim, S.; Javed, M. S.; Khan, J.; Khalid, M.; Shah, S. S. A.; Idrees, M.; Imran, M.; Usman, M.; Hussain, S.; Ahmad, I., Energy storage performance of binder-free ruthenium-oxide nano-needles based free-standing electrode in neutral pH electrolytes. Electrochimica Acta 2021, 378, 138139.
65. Duignan, T. T.; Zhao, X. S., The Born model can accurately describe electrostatic ion solvation. Physical Chemistry Chemical Physics 2020, 22 (43), 25126-25135.
66. Akkermans, R. L., Solvation Free Energy of Regular and Azeotropic Molecular Mixtures. The Journal of Physical Chemistry B 2017, 121 (7), 1675-1683.
67. Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G., Minnesota solvation database. Minnesota Solvation Database version 2012, 20.

Chapter 8

Publication: Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model

8.1 Scope of the Project

8.1.1 Project overview and motivation:

Continuum solvation models as one of the two major methods to take into account the solvent effects are extensively required in a wide range of scientific disciplines. The conventional continuum solvation models available for this purpose typically consider the solvation free energy as contributions from different energy attributes of the implicitly perturbed Hamiltonian and ad-hoc integration of them. Additionally, the current continuum solvation models employ conventional molecular surfaces which as we demonstrated in chapter 5, are not so efficient in describing the thermodynamics of the condensed phase. As a result, the conventional continuum solvation models commonly suffer from limited accuracy. With that motivation, in the current study, we investigated a more rigorous integration of the same energy attributes of the implicitly perturbed Hamiltonian via machine learning for achieving higher accuracy and reliability.

8.1.2 Novelty aspects:

- Introducing ML-PCM as a rigorous machine learning approach to integrate energy attributes in conventional continuum solvation models
- Improving the accuracy of conventional continuum solvation models by one order of magnitude
- Providing a highly accurate evaluation of solvation free energy with almost no additional computational cost compared to conventional models

8.1.3 Connection to other chapters:

The provided machine learning models were developed based on the guidelines proposed in chapter 6. The general ideal of this project was inspired by studying the implicit solvent approach in chapter 4. The molecular surfaces employed as inputs of the machine-learning model were in close connection to the conception of the thermodynamically effective surfaces in chapter 5. The newly developed method was also the cornerstone of conceiving the ImPerHam representations in chapter 9.

8.2 Publication Data and Reprint

Reference: Alibakshi, A., Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. *Nature Communications* **2021,** *12(1), pp.1-7*.

Submitted: 11 August 2020

Accepted: 12 May 2021

Contribution: Carrying out all the computations, method development, and writing the manuscript

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23724-6 OPEN

Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model

Amin Alibakhshi [1]✉ & Bernd Hartke [1]

Theoretical estimation of solvation free energy by continuum solvation models, as a standard approach in computational chemistry, is extensively applied by a broad range of scientific disciplines. Nevertheless, the current widely accepted solvation models are either inaccurate in reproducing experimentally determined solvation free energies or require a number of macroscopic observables which are not always readily available. In the present study, we develop and introduce the Machine-Learning Polarizable Continuum solvation Model (ML-PCM) for a substantial improvement of the predictability of solvation free energy. The performance and reliability of the developed models are validated through a rigorous and demanding validation procedure. The ML-PCM models developed in the present study improve the accuracy of widely accepted continuum solvation models by almost one order of magnitude with almost no additional computational costs. A freely available software is developed and provided for a straightforward implementation of the new approach.

[1] Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, Kiel, Germany. ✉email: alibakhshi@pctc.uni-kiel.de

Free energy of solvation is one of the key thermophysical properties in studying thermochemistry in solution, where the majority of real-life chemistry happens. In theoretical studies of solution chemistry, estimation of free energies allows evaluation of reaction rates and equilibrium constants of physical or chemical reactions of interest. Nevertheless, direct evaluation of free energies in solution can be quite challenging since it sometimes requires appropriate sampling of phase space[1–3] and appropriate treatment of the non-covalent interactions between the solvent and solute, which can have a remarkable impact on electronic structures of both the solvent and solute and consequently on the microscopic and macroscopic observables[4,5].

Theoretical approaches for evaluating physical chemistry behind solvation free energy can be generally divided into two main categories, namely explicit solvent and implicit solvent approaches. In explicit solvent approaches, solvent molecules are treated explicitly, and the free energy is typically evaluated by analyzing the trajectory of time evolution of phase space obtained via molecular dynamics or Monte Carlo simulations. For that end, a number of efficient free energy estimators have been developed in the past decades such as thermodynamic integration, free-energy perturbation, and histogram analysis methods[6].

Despite obvious advantages of applying the explicit solvent methods such as retaining the physically proper picture of discrete solvent molecules, they suffer by a number of limitations when applied to free-energy estimation. For example, in case of applying methods which evaluate the free energy through alchemical transformations (e.g., thermodynamic integration or free energy perturbation), defining intermediate states and pathways between the endpoints appropriately can be quite tricky[7]. Also, necessity of employing appropriate force fields, which for many solute-solvent mixtures requires to develop or reparametrize a force field, and running the simulations and trajectory analyses can be laborious and time-taking tasks.

To overcome the mentioned limitations, the implicit solvent approach has been developed and is widely applied as standard method for studying solvent effects in computational chemistry. In implicit solvent approaches, the solvent molecules are treated implicitly as a continuous medium and the solute is placed in a cavity of this implicitly defined solvent. The solute-solvent interactions are then evaluated via considering the solvent polarization due to the solute charge distribution and its resulting potential field acting on the solute, known as the reaction field[5]. For a moderate level of theory and medium-sized molecules, implicit solvent approaches can yield a reasonable estimation of the solvation free energy in few seconds to few minutes on a normal desktop PC, while for explicit solvent approaches it might take from hours to days.

The most widely applied implicit solvent approaches are those based on the so-called polarizable continuum model (PCM) proposed by Tomasi and co-workers[8]. In polarizable continuum models, the solvation free energy is constructed by summing the contributions of electrostatic interactions including electronic, nuclear, and polarization interactions (ΔG_{ENP}), changes in free energy by solvent cavity formation, dispersion energy and local solvent structure changes (G_{CDS}), and corrections for differences in molar densities in the two phases compared with the standard state (ΔG°_{cons}). The contributions of electrostatic interactions are evaluated by iteratively solving the following relationship:

$$\Delta G_{ENP} = \langle \Psi^{(1)}|H + \frac{1}{2}V|\Psi^{(1)}\rangle - \langle \Psi^{(0)}|H|\Psi^{(0)}\rangle \qquad (1)$$

which is known as the self-consistent reaction-field (SCRF) calculations[5]. Here, superscripts (0) and (1) refer to the gas and solution phases, respectively, and V is the potential energy operator resulting from the reaction field. Various constructions of the potential energy operator as well as G_{CDS} have resulted in different continuum solvation models. The parallel existence of several continuum solvation models is a good indicator that each of them has its own strengths and weaknesses, and choosing a single, optimal model is not trivial. It is totally impossible to provide a detailed overview here; a 2005 review of implicit solvation models[9] covered 95 pages and cited 936 references. In the present study, we only consider the most widely used PCM-based models.

One of simplest and yet successful continuum solvation models is CPCM which implements the conductor-like screening solvation boundary condition within the PCM framework. In CPCM, the following correction of the polarization charge densities by the scaling factor x is employed[10]:

$$f(\varepsilon) = \frac{\varepsilon - 1}{\varepsilon + x} \qquad (2)$$

where ε is the solvent dielectric constant. One main advantage of CPCM is its much simpler defined boundary conditions. More importantly, unlike more advanced PCM-based models which require the normal component of the solute electric field as input, CPCM only requires the solute electrostatic potential; for this reason it is much less affected by outlying charge errors (OCE)[11,12]. A more versatile model exploiting the conductor-like screening solvation boundary condition is COSMO-RS, developed by Klamt and co-workers[13,14], which although initially proposed in 1995, still is one of the most accurate available continuum solvation models. A more sophisticated treatment of the boundary condition is implemented in the integral equation formalism of PCM (IEF-PCM) taking into account apparent surface charge isotropic[15] or anisotropic[16] dielectric continuum solvation. Another extensively used continuum solvation model is the SMx family of methods which specifically focuses on more accurate estimation of the solvation free energy[4,5].

We already discussed the main advantages of continuum solvation models such as their efficiency in terms of computational cost. Nevertheless, it should be noted that all this has become possible for a considerable amount of assumptions and simplifications on the physics of the problem, such as overlooking the conformational entropy of solvent and solute which can have a significant contribution on the total free energy[17], neglecting the site-specific solute-solvent interactions and decoupling the polar and nonpolar components of free energies and considering them independent, linear and additive[18,19]. The inaccuracies resulting from such simplifications are commonly compensated for via incorporating additional macroscopic observables as well as adjustable parameters in the solvation models. In the CPCM model for example, this is achieved by implementing an ad hoc modification of the atomic radii via defining a number of adjustable parameters and empirical descriptors, such as the number of bonded hydrogens and the number of bonded active atoms[10]. In the COSMO-RS model, it is achieved by ad hoc modification of the interaction energies and effective contact area via some adjustable parameters[14].

In contrast, in the SMx family of methods, to provide a more accurate estimation of the solvation free energy, an ad hoc modification of the G_{CDS} term in (1) has been proposed. For that end, employing additional macroscopic observables in the model has been considered[4], including the refractive index, Abraham's hydrogen bond acidity and basicity of the solute, macroscopic surface tension of the solvent at the air/solvent interface at 298.15 K, the square of the fraction of solvent atoms that are aromatic carbon atoms, and the square of the fraction of solvent atoms that are F, Cl, or Br. Although these employed macroscopic observables indirectly introduce more physics into the model and hence provide the chance to make predictions of solvation free

energies more universal, except for the last two they are not readily available for many new compounds and their experimental or theoretical evaluation is not straightforward.

In a number of recent studies, Machine Learning (ML) has been exploited to map the highly complicated relationship between solvation free energy and potentially relevant macroscopic or microscopic observables.

Wang et al. employed a pool of 30 molecular representations which all are either per atom reaction field energies or partial charges, as the input of the learning-to rank (LTR) machine learning algorithm, resulting in a root mean squared error (RMSE) of 1.05 kcal/mol[18]. Borhani et al. developed a QSPR model which requires 12 experimentally determined properties of solvent and 9 QM derived representations of solute as model input, yielding a Mean Unsigned Error (MUE) of 0.43 kcal/mol[20]. Hutchinson and Kobayashi proposed a structure property relationship for prediction of hydration free energy which yields a RMSE of 1.65 kcal/mol[21].

Another recent example is the kernel-based machine learning model of Rauer and Bereau which is developed to predict the free energy of solvating small organic molecules containing C, H, O, and N atoms in pure water via implicit-solvent molecular dynamics simulations[22]. For a 39-parameter model they reported a MUE of 1.06 kcal/mol.

The most recent example of employing machine learning for prediction of solvation free energy is the model developed by Vermeire and Green[23]. Their model is developed based on the transfer of knowledge learned through one million data of QM evaluated free energies and fine tuning it to accurately reproduce the experimentally determined solvation free energies. They reported a MUE of 0.21 kcal/mol for their model which is currently the most accurate ever reported result for prediction of solvation free energy.

In the present study, we propose a machine-learning-based PCM model, which, similar to other conventional continuum solvation models, is based on considering the solvent as a continuous medium and calculating the solvation energy components of a solute placed in the cavity of this medium by the SCRF procedure. Nevertheless, unlike the conventional PCM models which propose simple and ad hoc expressions to integrate and modify those calculated energy components, we employ machine learning for this purpose and show its efficiency in substantial improvements of the predictability of solvation free energy.

Results and discussions

After setting up and training the neural networks and screening the appropriately trained models via the post-validation strategy discussed in the previous section, the best results with MUE of 0.52526 and 0.40011 kcal/mol were observed for the computations at B3LYP/6–31 G* and DSD-PBEP86-D3/def2TZVP levels of theory, respectively. The two models employed SCRF energy components and solvation free energy computed via $CPCM_{x=0.5}$ solvation model in both cases and 100 and 130 hidden layer neurons, respectively. These two models are denoted by ML-PCM(B3LYP) and ML-PCM (DSD-PBEP86) hereafter, respectively. Details of the selected input variables and implementation instructions for all selected models are provided in Supplementary Software 1. These results show a substantial improvement compared to the original continuum solvation model $CPCM_{x=0.5}$, which for the same dataset yielded MUE of 3.1611 and 2.9130 kcal/mol, respectively.

In comparison to the SMD model, which for the same dataset and solvation free energy computations at B3LYP/6–31 G* and DSD-PBEP86-D3/def2TZVP levels yields MUE of 0.78623 and 0.85396 kcal/mol, respectively, the obtained results still show a higher accuracy, without requiring additional solvent parameters needed in the SMD approach. In comparison to the MUE of 0.4214 kcal/mol reported by Klamt and Diedenhofen[24] for employing one of the recent versions of the COSMO-RS model for the same dataset, the ML-PCM(DSD-PBEP86) provides a slightly higher accuracy. Also, in terms of maximum unsigned error, the two ML-PCM models which yield maximum unsigned error of 6.2252and 3.8799 kcal/mol, respectively, are more accurate than that of COSMO-RS for which this value is 6.8701 kcal/mol. For other continuum solvation models studied for the same dataset, the maximum unsigned error of the SMD, PCM, CPCM and $CPCM_{x=0.5}$ were 11.311, 12.75, 12.2, 12.6 kcal/mol for B3LYP/6–31 G* and 11.311, 12.83, 12.31, 12.68 kcal/mol for DSD-PBEP86-D3/def2TZVP levels of theory, which are all substantially higher than those achievable by the ML based models.

The higher accuracy of the predicted solvation free energies by the COSMO-RS model compared to the other conventional solvation models also motivated us to study neural networks which take SCRF energy components computed via PCM or CPCM models in addition to the solvation free energies predicted via COSMO-RS as neural network feeds. For these updates, the best results with MUEs of 0.26057 and 0.24387 kcal/mol and maximum unsigned errors of 7.1349 and 2.9154 kcal/mol were obtained for energy components calculated via $CPCM_{x=0.5}$ and CPCM solvation models, 130 and 120 hidden layer neurons, and computations at B3LYP/6–31 G* and DSD-PBEP86-D3/def2TZVP levels of theory, respectively. These two models, which are denoted by ML-PCM/COSMO-RS(B3LYP) and ML-PCM/COSMO-RS(DSD-PBEP86) hereafter, respectively, show a remarkable improvement in predicted solvation free energy compared to those obtained via the original implementation of COSMO-RS reported by Klamt and Diedenhofen[24]. This implies considerable flexibility of the proposed approach in improving accuracy of various solvation models. Nevertheless, it should be noted that the solvation free energies evaluated by COSMO-RS which were used as additional model inputs in the present study were evaluated using the 2015 version of that method. Using free energies evaluated by more recent versions of COSMO-RS and also the energy terms computed with this method, will probably result in more accurate predictions of the solvation free energy by the presented ML-PCM.

As the most important parameter in developing ANN models, we studied the impact of the selected number of hidden layer neurons on the performance of the developed machine learning models. As can be seen in Fig. 1, by increasing the number of hidden layer neurons, the predictability of the solvation free

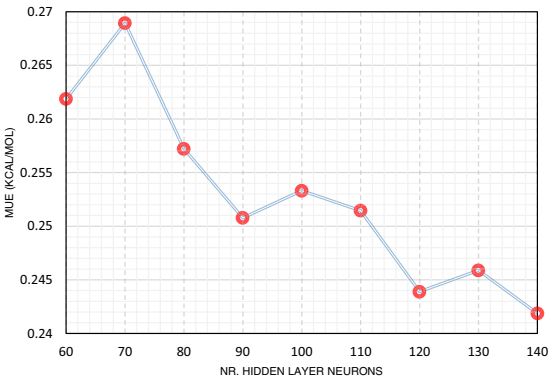


Fig. 1 MUE of developed ML-PCM/COSMO-RS(B3LYP) models versus the number of hidden layer neurons. The general trend shows the reducing pattern in MUE with increasing the size of the neural network model.

energy is generally improved. This is due to the larger number of adjustable parameters of the resulting models and their consequently higher flexibility to map complicated functionalities. However, at the same time this may reduce the extrapolation capability of the model, i.e., it may reduce performance when applied to samples remarkably different from those already examined in developing the models.

To investigate the impact of the number of hidden layer neurons on extrapolation performance of the models developed in the present study, we re-examined the trained models for out-of-sample predictions, following the approach proposed by Vermeire and Green[23]. For that end, we compared the results of models for which a group of samples with either a specific element or a specific solvent were included in the training dataset with the same models trained with a dataset excluding that specific group of samples. We studied out-of-sample prediction performance for 20 solvents and 6 solute elements most frequently encountered in our studied dataset. The obtained results are reported in Tables 1 and 2. According to the results, the developed models show an excellent extrapolation capability for out of sample predictions of solvent splits, while for the element splits, the extrapolation is slightly less accurate. Furthermore, except for the element Br, the out-of-sample predictions tested for ML-PCM/COSMO-RS(B3LYP) are within chemical accuracy.

A comparison of predicted and experimentally determined free energies is depicted in Fig. 2. As can be seen, the linear correlation between the predicted and reference data is more evident for the newly derived models, compared to the conventionally accepted ones.

The overall results obtained via newly developed ML models are compared with various other models proposed in the literature in Table 3. Although a more informative comparison would be possible if different models were compared for the same dataset and, if applicable, the same level of theory, the larger size of the benchmark dataset used in the present study compared to most of the other works confirms the superior accuracy of the newly proposed method compared to the majority of the widely accepted ones. In comparison to the model developed by Vermeire and Green[23] which yields MUE of 0.21 kcal/mol, our results are slightly less accurate, but it should be noted that our results are obtained for a much lower number of neurons and model parameters.

Furthermore, it should be noted that the inaccuracies inherent in the reference data of solvation free energies (Aleatoric uncertainty) can also impact both the training efficiency and inferences about model performances, as pointed out by Vermeire and Green[23].

To summarize, we have demonstrated substantial improvements of continuum solvation models in evaluating solvation free energy with the help of machine learning. For that end, we proposed a more versatile machine learning assisted integration of the continuum solvation energy components calculated in SCRF computations which can be used to modify the predicted solvation free energy by various solvation models. It allowed us to achieve

Table 1 Out-of-sample predictions for solvent splits.

		ML-PCM/COSMO-RS(B3LYP)		ML-PCM/COSMO-RS(DSD-PBEP86)	
Solvent	**Nr. Samples**	**MUE (solvent included)**	**MUE (solvent excluded)**	**MUE (solvent included)**	**MUE (solvent excluded)**
Water	261	0.13921	0.53856	0.12724	0.52107
n-Octanol	199	0.21116	0.40528	0.19416	0.34079
n-Hexadecane	184	0.47931	0.63312	0.42914	0.38652
Chloroform	102	0.2962	0.33126	0.27975	0.28894
CycloHexane	88	0.27941	0.30729	0.30521	0.35877
CarbonTetraChloride	73	0.37704	0.38958	0.30407	0.31146
Benzene	71	0.21953	0.24581	0.37323	0.52627
DiethylEther	66	0.23975	0.29187	0.22181	0.22156
Heptane	64	0.41215	0.4795	0.2033	0.19233
n-Hexane	57	0.19548	0.19648	0.28332	0.3775
Toluene	49	0.22219	0.20023	0.31435	0.33986
Xylene-mixture	46	0.25694	0.22309	0.27209	0.27789
DiChloroEthane	37	0.38469	0.49085	0.22075	0.28748
n-Decane	37	0.21171	0.25761	0.15559	0.17148
ChloroBenzene	36	0.2183	0.2374	0.20119	0.22425
n-Octane	35	0.13265	0.13455	0.17431	0.19447
2,2,4-TriMethylPentane	32	0.2097	0.20388	0.23426	0.23618
EthylBenzene	27	0.20878	0.23166	0.20471	0.24331
BromoBenzene	24	0.16054	0.22188	0.13648	0.18478
Decalin-mixture	24	0.39408	0.44484	0.31164	0.33004

Table 2 Out-of-sample predictions for element splits.

		ML-PCM/COSMO-RS(B3LYP)		ML-PCM/COSMO-RS(DSD-PBEP86)	
Element	**Nr. Samples**	**MUE (element included)**	**MUE (element excluded)**	**MUE (element included)**	**MUE (element excluded)**
N	611	0.25549	0.40139	0.25486	0.37139
F	81	0.29188	0.38812	0.32345	0.48087
P	62	0.12927	0.64773	0.20648	0.95936
S	91	0.26592	0.50868	0.29104	0.53079
Cl	174	0.25295	0.5194	0.17956	0.47383
Br	102	0.25005	1.4559	0.26268	0.91972

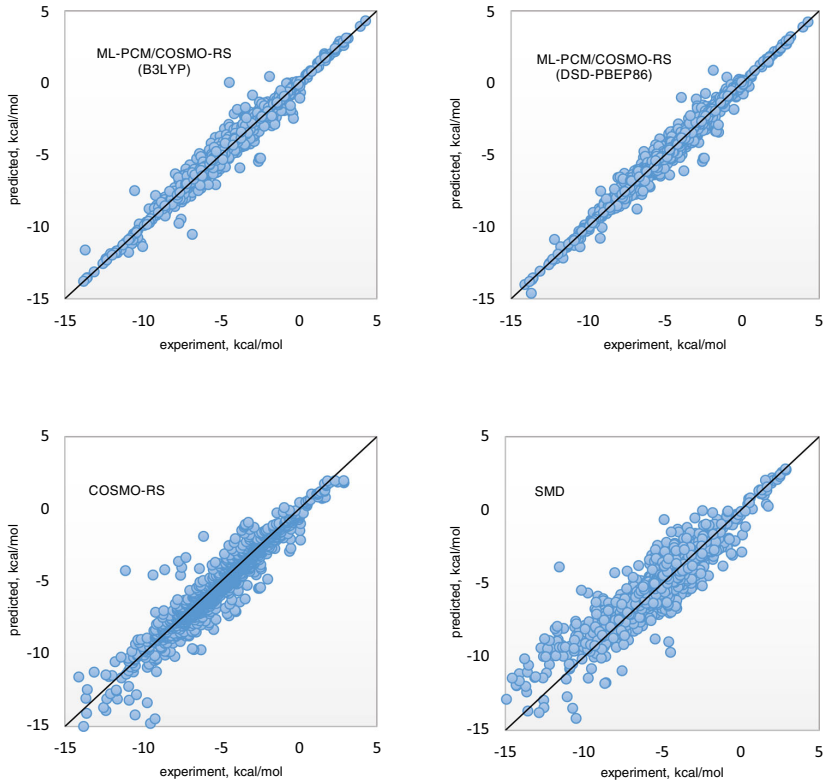


Fig. 2 Comparison of experimentally determined and predicted solvation free energies for various solvation models. The results show a higher correlation between the experimentally determined and predicted data for the proposed machine learning solvation models compared to the SMD or COSMO-RS models.

accurate predictions of solvation free energy with MUE as low as 0.2439 kcal/mol for a large dataset of 2493 binary mixtures of 435 neutral solutes and 91 solvents from diverse chemical families.

Methods

Dataset. To benchmark our results, we used the solvation free energy data of 2493 binary mixtures of 435 neutral solutes and 91 solvents from diverse chemical families available in the Minnesota solvation database[4]. The full list of the studied samples can be found as Supplementary Data 1.

Computational details. The performance of models is reported as mean unsigned error (MUE) and root mean squared error (RMSE) defined as:

$$MUE = \frac{1}{N}\sum\left(\left|y_i^{\exp} - y_i^{pred}\right|\right) \quad (3)$$

$$RMSE = \left(\frac{1}{N}\sum\left(\left(y_i^{\exp} - y_i^{pred}\right)^2\right)\right)^{\frac{1}{2}} \quad (4)$$

where $y_i^{\exp}$ and y_i^{pred} are experimentally determined and predicted solvation free energies, respectively.

Prior to SCRF computations, all solute geometries were optimized in vacuo at the B3LYP/6-31 G*level of theory. Using the optimized structures, the SCRF principal energy components listed in Table 4 were computed for each compound at the B3LYP/6-31 G* and DSD-PBEP86-D3/def2TZVP levels of theory. The latter method as a double hybrid has been shown to yield more precise charge distributions and energy estimations compared to lower-rung DFT or MP2 methods, for a cost comparable to that of the MP2 calculation[25].

The SCRF energy components listed in Table 4 were computed for two widely accepted polarizable continuum models, namely the IEF-PCM and CPCM, as implemented in Gaussian 16 (ref. [26]). For CPCM, the default value of zero is considered as the scaling factor x in relationship (2). However, a value of 0.5 has been shown to be a more reasonable choice for this scaling factor[11,27]. Therefore, in addition to the default implementation of CPCM in Gaussian 16, we also employed a CPCM model with a scaling factor of x=0.5 and denote it by $CPCM_{x=0.5}$. For that, we replaced the original dielectric constant of the solvent with an effective dielectric constant $\tilde{\varepsilon}(\varepsilon, x)$ calculated via:

$$\tilde{\varepsilon}(\varepsilon, x) = \frac{\varepsilon + x}{x + 1} \quad (5)$$

as suggested by Klamt et al.[11]. For comparison purposes, we also calculated the solvation free energy via the SMD approach.

We employed feed-forward neural networks to map the relationship between the solvation free energy and the calculated SCRF energy components, which in addition to the solvation free energy estimated by the applied continuum solvation model and to the dielectric constant of the solvent, comprised our model inputs.

The obtained pool of model inputs was further screened using the Minimum Redundancy and Maximum Relevance (MRMR) algorithm[28] resulting in various 8–16 membered combinations of those variables. MRMR is a highly efficient algorithms for selecting most effective sets of variables for developing robust machine-learning-based models[29]. For each number of selected variables, 25 different settings of the MRMR algorithm were applied, distinguished by the employed quantization level, level of dependency, forward or backward variable selection and considering pseudo-samples based on Bayesian statistics or not[28]. In many cases, this resulted in diversely selected set of variables, even for the same applied level of theory and continuum solvation model.

In the next step, various configurations of neural network models were set up and their reliability were examined with a demanding procedure based on the guidelines presented in a previous study[30]. Accordingly, we assigned large parts of the dataset for test (25%) and validation (15%), and only 60% of the dataset compounds were used for training the models.

To improve the transferability of the developed models for out-of-sample predictions, validation and test sets were selected in a way to include either solvent or solute elements not available in the training set.

We employed Levenberg-Marquardt backpropagation and Gradient descent backpropagation training algorithms, and hidden layer transfer functions of the logarithm-sigmoid and tangent-sigmoid types[31]. We only employed neural

Table 3 Comparison of the results of the new method with other models.

Method	Source	Nr. Samples	Nr. Solvents	Nr. Solutes	Deviation measure	Deviation (kcal/mol)
ML-PCM/COSMO-RS(DSD-PBEP86)	Present study	2224	88	300	MUE	0.24387
					RMSE	0.37252
ML-PCM/COSMO-RS(B3LYP)	Present study	2224	88	300	MUE	0.26057
					RMSE	0.43623
ML-PCM (DSD-PBEP86)	Present study	2488	91	435	MUE	0.40011
					RMSE	0.56014
ML-PCM (B3LYP)	Present study	2493	91	435	MUE	0.52526
					RMSE	0.75112
Other models found in the literature:						
Machine learning	Vermeire and Green[23]	10145	291	1368	MUE	0.21
					RMSE	0.44
COSMO-RS	Klamt and Diedenhofen[24]	2346	91	318	MUE	0.42145
					RMSE	0.69644
SM12	Marenich et al.[32]	2403	91	352	MUE	0.5457-0.6717
QSPR	Borhani et al.[20]	1777	210	295	MUE	0.43
					RMSE	0.52
DCOSMO-RS	Klamt and Diedenhofen[24]	2346	91	318	MUE	0.6584
					RMSE	0.99724
SMD (B3LYP)	Present study	2493	91	435	MUE	0.78623
					RMSE	1.1633
SMD (DSD-PBEP86)	Present study	2488	91	435	MUE	0.85396
					RMSE	1.3362
Feature Functional Theory	Wang et al.[18]	668	1 (water)	668	RMSE	1.05
kernel-based machine learning	Rauer and Bereau[22]	355	1 (water)	355	MUE	1.06
atoms-in-molecules neural network	Zubatyuk et al.[33]	-	-	414	MUE	1.1
Structure-Property Relationship	Hutchinson and Kobayashi[21]	-	1 (water)	-	RMSE	1.65
CPCM(B3LYP)	Present study	2493	91	435	MUE	2.6942
					RMSE	3.1733
PCM(B3LYP)	Present study	2493	91	435	MUE	2.9054
					RMSE	3.3948
$CPCM_{x=0.5}$(B3LYP)	Present study	2493	91	435	MUE	2.9130
					RMSE	3.3985
CPCM(DSD-PBEP86)	Present study	2488	91	435	MUE	2.9651
					RMSE	3.4426
PCM (DSD-PBEP86)	Present study	2488	91	435	MUE	3.1569
					RMSE	3.6445
$CPCM_{x=0.5}$(DSD-PBEP86)	Present study	2488	91	435	MUE	3.1611
					RMSE	3.6466

Table 4 The components of the continuum solvation model.

1	Solvation free energy calculated by the continuum solvation model
2	$\langle\Psi^{(0)}\|H\|\Psi^{(0)}\rangle$
3	$\langle\Psi^{(0)}\|H+V^{(0)}/2\|\Psi^{(0)}\rangle$
4	$\langle\Psi^{(0)}\|H+V^{(1)}/2\|\Psi^{(0)}\rangle$
5	$\langle\Psi^{(1)}\|H\|\Psi^{(1)}\rangle$
6	$\langle\Psi^{(1)}\|H+V^{(1)}/2\|\Psi^{(1)}\rangle$
7	Interaction energy of unpolarized solute and polarized solvent
8	Interaction energy of polarized solute and polarized solvent
9	Solute polarization energy
10	Total electrostatic interaction energy
11	Cavity surface area
12	Cavity volume
13	Total kinetic energy
14	Total potential energy
15	Sum of kinetic and potential energy

networks with one hidden layer and 1 to 140 neurons in the hidden layer, with intervals of 10 neurons for ANNs with more than 50 neurons in the hidden layer. For each neural network configuration, training was carried out for 60 randomly selected training, validation and test sets, and for each one 40 different initializations of weight and bias constants of the neural networks were made. Above all, to avoid getting misleading data affected by favorable or unfavorable division of dataset into training, validation and test sets, the post validation strategy proposed in a previous study[30] was carried out. Accordingly, during the initial training of the neural networks, for the models which yielded mean absolute percentage errors lower than 22%, the final optimized weights and bias constants of the neural network models were recorded. These recorded constants were used as the initial guess to train, validate and test the same neural network configurations but under 100 different randomly selected training, validation and test sets. The models for which in at least 80 out of 100 iterations their test and training sets errors had the same means and variances as evaluated by the two sample t-test method with 5% significance level were considered as reliably trained models. For them, the average of the ANN-predicted results in all repeats were reported as the performance of that model. Setting up and running the neural network models were implemented in Matlab software. A freely available C++ code for practical use of our proposed ML-PCM models, with detailed user instructions, is provided in Supplementary Software 1.

All the computations were carried out on the High Performance Computing center clusters of the Christian-Albrechts-University of Kiel.

Data availability

All data produced in this study are available and can be provided by contacting the corresponding author.

Code availability

The source file of the C++ code developed for implementing the proposed method with detailed used instructions are available in Supplementary Software 1 or can be provided by contacting the corresponding author.

Received: 11 August 2020; Accepted: 12 May 2021;
Published online: 18 June 2021

References

1. Dittner, M. & Hartke, B. Globally optimal catalytic fields–inverse design of abstract embeddings for maximum reaction rate acceleration. *J. Chem. theory Comput.* **14**, 3547–3564 (2018).
2. Gauthier, J. A., Dickens, C. F., Chen, L. D., Doyle, A. D. & Nørskov, J. K. Solvation effects for oxygen evolution reaction catalysis on IrO2 (110). The. *J. Phys. Chem. C.* **121**, 11455–11463 (2017).
3. Sakong, S. & Groß, A. The importance of the electrochemical environment in the electro-oxidation of methanol on Pt (111). *ACS Catal.* **6**, 5575–5586 (2016).

4. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).
5. Cramer, C. J. & Truhlar, D. G. A universal approach to solvation modeling. *Acc. Chem. Res.* **41**, 760–768 (2008).
6. Chipot, C. & Pohorille, A. *Free energy calculations*. (Springer, 2007).
7. Pohorille, A., Jarzynski, C. & Chipot, C. Good practices in free-energy calculations. *J. Phys. Chem. B* **114**, 10235–10253 (2010).
8. Miertuš, S., Scrocco, E. & Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **55**, 117–129 (1981).
9. Tomasi, J., Mennucci, B. & Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **105**, 2999 (2005).
10. Barone, V. & Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **102**, 1995–2001 (1998).
11. Klamt, A., Moya, C. & Palomar, J. A comprehensive comparison of the IEFPCM and SS (V) PE continuum solvation methods with the COSMO approach. *J. Chem. theory Comput.* **11**, 4220–4225 (2015).
12. Klamt, A. & Jonas, V. Treatment of the outlying charge in continuum solvation models. *J. Chem. Phys.* **105**, 9972–9981 (1996).
13. Klamt, A. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **99**, 2224–2235 (1995).
14. Klamt, A., Jonas, V., Bürger, T. & Lohrenz, J. C. Refinement and parametrization of COSMO-RS. The. *J. Phys. Chem. A* **102**, 5074–5085 (1998).
15. Mennucci, B., Cammi, R. & Tomasi, J. Excited states and solvatochromic shifts within a nonequilibrium solvation approach: A new formulation of the integral equation formalism method at the self-consistent field, configuration interaction, and multiconfiguration self-consistent field level. *J. Chem. Phys.* **109**, 2798–2807 (1998).
16. Cances, E., Mennucci, B. & Tomasi, J. A new integral equation formalism for the polarizable continuum model: theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **107**, 3032–3041 (1997).
17. Suárez, E., Díaz, N. & Suárez, D. Entropy calculations of single molecules by combining the rigid–rotor and harmonic-oscillator approximations with conformational entropy estimations from molecular dynamics simulations. *J. Chem. Theory Comput.* **7**, 2638–2653 (2011).
18. Wang, B., Wang, C., Wu, K. & Wei, G. W. Breaking the polar-nonpolar division in solvation free energy prediction. *J. Comput. Chem.* **39**, 217–233 (2018).
19. Dzubiella, J., Swanson, J. M. & McCammon, J. Coupling hydrophobicity, dispersion, and electrostatics in continuum solvent models. *Phys. Rev. Lett.* **96**, 087802 (2006).
20. Borhani, T. N., García-Muñoz, S., Luciani, C. V., Galindo, A. & Adjiman, C. S. Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Phys. Chem. Chem. Phys.* **21**, 13706–13720 (2019).
21. Hutchinson, S. T. & Kobayashi, R. Solvent-specific featurization for predicting free energies of solvation through machine learning. *J. Chem. Inf. Modeling* **59**, 1338–1346 (2019).
22. Rauer, C. & Bereau, T. Hydration free energies from kernel-based machine learning: compound-database bias. *J. Chem. Phys.* **153**, 014101 (2020).
23. Vermeire, F. H. & Green, W. H. Transfer learning for solvation free energies: from quantum chemistry to experiments. *Chem. Eng. J.* **418**, 129307 (2021).
24. Klamt, A. & Diedenhofen, M. Calculation of solvation free energies with DCOSMO-RS. *J. Phys. Chem. A* **119**, 5439–5445 (2015).
25. Kozuch, S. & Martin, J. M. DSD-PBEP86: in search of the best double-hybrid DFT with spin-component scaled MP2 and dispersion corrections. *Phys. Chem. Chem. Phys.* **13**, 20104–20107 (2011).
26. Frisch, M. et al. (Gaussian, Inc. Wallingford, CT, 2016).
27. Cossi, M., Rega, N., Scalmani, G. & Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **24**, 669–681 (2003).
28. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).
29. Brown, Gavin. A new perspective for information theoretic feature selection. Artificial intelligence and statistics, PMLR, pp. 49–56 (2009).
30. Alibakshi, A. Strategies to develop robust neural network models: prediction of flash point as a case study. *Anal. Chim. Acta* **1026**, 69–76 (2018).
31. Demuth, H. & Beale, M. Neural Network Toolbox For Use with Matlab--User'S Guide Verion 3.0. (1993).
32. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Generalized born solvation model SM12. *J. Chem. Theory Comput.* **9**, 609–620 (2013).
33. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **5**, eaav6490 (2019).

Acknowledgements

The authors wish to thank Karsten Balzer at the high performance computing center of Kiel University for his support and assistance in running the computations there. The Authors wish to thank the referees for their careful review of our work and fruitful discussions and comments.

Author contributions

A.A. has contributed to method development, carried out the computations and contributed to writing the manuscript. B.H. supervised the project and contributed to method development and writing the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23724-6.

Correspondence and requests for materials should be addressed to A.A.

Peer review information *Nature Communications* thanks John Keith and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 9

Implicitly perturbed Hamiltonian: a new class of molecular representations for machine learning

Project overview and motivation:

Molecular representations play the central role in the machine-learning study of molecular sciences. The most extensively applied molecular representations currently available, suffer from a number of limitations such as not being transferrable, limited accuracy and domain of applications, and non-physical functional form. With that motivation and inspired by the machine-learning model introduced in chapter 8, in this project we propsoed Implicitly Perturbed Hamiltonian (ImPerHam) as a new class of molecular representations to address the limitations of the conventional representations. We studied machine-learning models based on ImPerHam for three challenging and diverse case studies namely predicting inhibition of the CYP450 enzyme, high precision and transferrable evaluation of configurational energy of molecular systems and accurately reproducing solvation free energies.

Novelty aspects:

- The newly defined molecular representations
- Providing a machine-learning model for reliably predicting inhibition of the CYP450 as an important consideration in drug design

- Providing machine-learning models for accurately estimating solvation free energies with significantly higher accuracies and lower computational cost compared to the conventional methods
- Providing machine-learning models for improving the accuracy of GFN2-xTB method in the evaluation of molecular energies by one order of magnitude

Connection to other chapters:

The provided machine-learning models were developed based on the guidelines proposed in chapter 6. The general idea of ImPerHam representations was inspired by the method introduced in chapter 8 and the knowledge acquired by studying implicit solvent models in chapter 4.

Contributions:

Conceiving the idea of ImPerHam representations, major contribution in method development, carrying out all the computations, major contribution in writing the manuscript.

Publication status:

This work is in the submission process and its preprint is available online (DOI:10.33774/chemrxiv-2021-6kqbw).

Implicitly perturbed Hamiltonian: a class of versatile and general-purpose molecular representations for machine learning

Amin Alibakhshi[1,*], Bernd Hartke[1]
Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, 24118 Kiel, Germany
Corresponding author: alibakhshi@pctc.uni-kiel.de

Abstract

Unraveling challenging problems by machine learning has recently become a hot topic in many scientific disciplines. For developing rigorous machine-learning models to study problems of interest in molecular sciences, translating molecular structures to quantitative representations as suitable machine-learning inputs play a central role. Many different molecular representations and the state-of-the-art ones, although efficient in studying numerous molecular features, still are suboptimal in many challenging cases, as discussed in the context of the present research. The main aim of the present study is to introduce the Implicitly Perturbed Hamiltonian (ImPerHam) as a class of versatile representations for more efficient machine learning of challenging problems in molecular sciences. ImPerHam representations are defined as energy attributes of the molecular Hamiltonian, implicitly perturbed by a number of hypothetic or real arbitrary solvents based on continuum solvation models. We demonstrate the outstanding performance of machine-learning models based on ImPerHam representations for three diverse and challenging cases of predicting inhibition of the CYP450 enzyme, high precision, and transferrable evaluation of conformational energy of molecular systems, and accurately reproducing solvation free energies for large benchmark sets.

9.1 Introduction

Employing machine learning for studying complicated scientific challenges has recently become a widely accepted approach in science. As a continuously growing field, machine learning has become a promising tool in studying many diverse areas in molecular sciences, ranging from major topics in life science such as synthetic biology [1], genomics [2], drug discovery [3], and cell biology [4], to major sub-fields of chemistry such as theoretical [5], organic [6], quantum [7,8], polymer [9] and synthetic [10] chemistry.

Despite this diversity in applications, at the heart of all machine-learning approaches in molecular sciences is a reliance on translating molecular structures to quantities understandable by the machine-learning process. These quantities, which are commonly known as molecular representations in computational chemistry and molecular descriptors in chem-informatics, are uniquely defined and are considered as molecular fingerprints. With the molecular representations defined, the main role of machine learning is then to learn the relationship between those representations and the properties of interest.

The earliest example of employing molecular representations for estimating materials properties, to our knowledge, is the group contribution method proposed in the middle of the last century [11]. This method considers a linear or non-linear dependency between molecular properties and the functional groups present in molecules. The success of this simply defined representation in predicting many properties of chemicals for several decades [12-19] has motivated its employment for approximating potential energies of molecular ensembles [20-23], as one of the most extensively studied and, at the same time, most challenging applications of machine learning in theoretical and quantum chemistry [24,25]. Nevertheless, achieving high accuracy in challenging problems like predicting of conformational energy of molecular systems typically requires more versatile molecular representations. To that end, a number of efficient representations have been proposed and widely used in recent years, such as atom-centered symmetry functions [26], the bispectrum of the neighbor density [27], smooth overlap of atomic positions [28], and the Coulomb matrix [29]. The popularity of these representations mainly stems from their remarkably fast and straightforward acquisition, typically requiring only elementary calculations on the geometrical data of the molecules.

These more advanced representations, despite being more efficient compared to elementary representations like functional group numbers and types, still suffer from the limited utility in studying many challenging and complicated problems of interest.
For machine-learning evaluation of conformational energies via those representations, despite several decades of progress, achieving quantum-chemical accuracy still has typically remained limited to simple molecular systems consisting of very few atom types [26,30-34]. As the other limitation, the application of machine-learning models based on employing the commonly applied representations is usually restricted only to the specific systems for which they were developed. This makes it necessary to re-train those machine-learning models for new systems, which is a highly demanding task. Finally, the currently defined representations mainly follow *non-physical* functional forms which result in very limited extrapolation capabilities [35].

For machine-learning evaluation of solvation free energy via conventional representations, in addition to the abovementioned limitations, the results obtainable via most of the currently developed models are still beyond chemical accuracy, as recently reviewed by us [36].

All these issues imply the necessity and importance of developing novel and more efficient representations, capable of addressing the above-mentioned shortcomings. For that purpose, appropriate representations have to satisfy a number of prerequisites such as invariance with respect to translation or rotation of the origin of the coordinate system, invariance to permutation of atoms of the same element, and yielding a unique or constant number of quantities independent of number and type of atoms in the system [34]. Ideally, representations should also allow transferability of machine-learning models to new systems, generate a large number of quantities to enhance machine-learning efficiency for diverse complicated properties of interest, and be physically interpretable.

With all these considerations, the main aim of the present study is to introduce the Implicitly Perturbed Hamiltonian (ImPerHam) as a new class of molecular representations to fulfill all the above-mentioned requirements.

The general idea behind ImPerHam originates from our recently developed machine-learning model for the evaluation of solvation free energy [37]. In that work, we employed machine learning for a more rigorous integration of the continuum solvation energy components. To that end, in addition to cavity geometrical data, we also employed energy attributes of Hamiltonian perturbed by the implicitly defined solvent as inputs of machine-learning models.

The outstanding efficiency of implicitly perturbed Hamiltonian energy attributes in characterizing the highly challenging case of solvation free energy reported in our recent study [37] motivated us to employ similar quantities computed not only for a single solvent of interest but also for a diverse set of other implicitly defined solvents and use them as molecular representations for other purposes as well.

To clarify the general idea behind the ImPerHam representations, we discuss the evaluation of solvation free energy of methane in water as an example. For that purpose, the conventional approaches require computing energy attributes of the methane Hamiltonian implicitly perturbed in water and integrating them based on the conventional approaches in continuum solvation models or via our recently proposed machine-learning model [37]. Based on the new approach, however, we still employ those energy attributes but computed for methane dissolved a number of other arbitrary real or hypothetic solvents and use them as molecular representations for methane.

An obvious advantage of ImPerHam representations stems from the provided possibility of generating an unlimited number of representations via an unlimited choice of solvents.

More importantly, the generated representations are actually thermodynamic quantities attributed to energy and free energy. Considering that based on the foundations of thermodynamics, the majority of system properties and thermodynamic quantities can be expressed as functions of energy and free energy, the representations based on these energy functions can be theoretically related to most properties of interest. For this reason, ImPerHam representations can be expected to perform better than the traditional representations mentioned above, which have less direct and less clear-cut relations to the properties of interest. As another direct result, although an unlimited number of solvents can be defined and employed to generate ImPerHam representations, as will be shown later in the present study, even very few arbitrary but diversely defined solvents can yield efficient machine-learning models for the diverse challenging problems considered in the present study.

The ImPerHam representations can be generated and studied using low-cost quantum mechanical computations or conventional polarizable force fields, which allows fast evaluation of those representations. Nevertheless, the computational cost of generating the ImPerHam representations can be somewhat higher than for other conventional representations discussed earlier. Despite this higher computational cost of computing ImPerHam representations, considering that the developed models based on them are highly transferable and do not require demanding re-training for new systems, we can still consider them an economic and reasonable choice.

In the present study, we demonstrate the efficiency of the machine-learning models based on ImPerHam representations in studying three challenging and extensively required problems in molecular sciences.

The first case study is the prediction of inhibition of a cytochrome P450 (CYP450) enzyme by small molecules. Prediction of CYP450 inhibition by potential drug candidates is a mandatory consideration in drug discovery, due to the crucially important role of these enzymes in metabolizing the majority of the drugs currently found in the market [38]. Inhibition of CYP450 by potential drugs can result in their accumulation in the body and increase the risk of drug-drug interactions [39]. Mibefradil and Cerivastatin are two examples of commercial drugs which have previously retracted from the market for this reason [40,41]. On the other hand, targeted inhibition of specific cytochromes has been proposed as an effective treatment strategy, as has been demonstrated for metastatic prostate tumors [42]. In the present study, we develop machine-learning models to predict inhibition of CYP450 1A2 as one of the important CYP450 members present in the liver. The structure of human microsomal CYP450 1A2 obtained by Sansen et al. [43] is depicted in figure 9.1.

In addition to inhibition of a CYP450 enzyme by machine-learning, we also investigate the performance of ImPerHam representations in the machine-learning evaluation of solvation

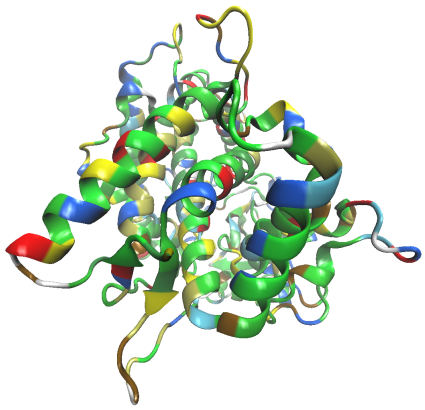


Fig. 9.1 Structure of human microsomal CYP450 1A2 enzyme.

free energy as well as conformational energy, benchmarked for large datasets. These case studies are not only among the most challenging problems in computational chemistry but also are extensively required in a wide range of scientific fields which makes them ideal for benchmarking the new representations.

9.2 Computational details

9.2.1 Benchmark sets

To be able to make a reliable evaluation of the newly proposed representations, we employed large benchmark sets commonly applied for each studied application.
Accordingly, for studying the inhibition of the CYP450 enzyme we exploited the benchmark set provided by Novotarskyi and co-workers for comparing different machine-learning models proposed for the same purpose [44]. We used the same compounds considered by them as training and test datasets to train and validate our model, which included 3745 and 3740 samples, respectively. Note that this benchmark data re-use implies that we have not done any molecule-protein docking calculations ourselves, but instead employed machine learning based on our ImPerHam representations to reproduce the inhibitor/non-inhibitor classification contained in this benchmark set (cf. subsections 9.2.2 and 9.2.3).

To benchmark our newly proposed representations in reproducing high-quality molecular energies, we exploited CCSD(T) reference energies computed for multiple configurations of 3525 dimers provided in the DES370K database [45]. To reduce the computational cost of model development, we randomly selected roughly half of the available dimers a priori, which yielded a dataset containing 1155 neutral, 390 positively charged, and 177 negatively charged dimers. From the available conformers of each one of the selected dimers, five were

selected for developing the machine-learning models, linearly distributed between the highest and lowest energy conformers.

For studying the predictability of solvation free energies, the efficiency of the developed machine-learning models was benchmarked using the Minnesota solvation database [46] containing 2493 reference solvation free energy data for binary mixtures of 435 solutes and 91 solvents.

The accuracy of the developed models is reported as the mean unsigned error (MUE) in kcal/mol, defined as:

$$\text{MUE} = \frac{1}{N} \sum \left(\left| y_i^{exp} - y_i^{pred} \right| \right). \tag{9.1}$$

9.2.2 Generating representations

We generated ImPerHam representations by perturbing the Hamiltonian via the analytical linearized Poisson-Boltzmann (ALPB) solvation model [47], as implemented in Grimme's Semiempirical Extended Tight-Binding Program Package [48] (xTB). The quantum-mechanical computations were carried out with the GFN2-xTB semiempirical method [49], for fast and computationally inexpensive acquisition and reasonable estimation of the studied energy attributes.

For each molecule in the vacuum state as well as in implicitly defined acetone, acetonitrile, benzene, hexane, water, ether, ethyl acetate, 1-octanol, and phenol solvents, the Hamiltonian energy attributes listed in table 9.1, were computed with original parameterizations as provided in xTB [48]. This constituted the initial pool of representations.

These solvents were selected from a wider choice of available solvents parameterized in xTB, with the intention to span a diverse range of dielectric constants. Nevertheless, any real or hypothetic solvent with an arbitrary or actual dielectric constant can also be used for this purpose, resulting in a wide range of relevant representations and higher flexibility of the machine-learning models to evaluate different quantities. This limited number of solvents was considered only to present the approach and efficiency of the proposed representations and as proof of concept.

For the computed Hamiltonians, the energy attributes reported in table 9.1 were considered as potentially relevant representations. Nevertheless, many other relevant energy attributes can also be defined and employed for this purpose.

For the machine-learning models developed to approximate conformational energies via ImPerHam representations, among the 5 selected configurations of each individual dimer, the conformation with the lowest energy was considered as the reference and its

Table 9.1 List of energy components of the implicitly perturbed Hamiltonian employed as molecular representations.

1	Total energy
2	Total free energy without cavity and hydrogen bonding contributions
3	HOMO-LUMO gap
4	HOMO orbital eigenvalue
5	LUMO orbital eigenvalue
6	SCC energy
7	Gibbs free energy (total)
8	Gibbs free energy (electric)
9	Gibbs free energy (cavitation)
10	Gibbs free energy (hydrogen bond)
11	Free energy shift for infinite dilution
12	Gradient norm
13	Isotropic electrostatic energy
14	Anisotropic exchange correlation
15	Isotropic exchange correlation
16	Dispersion energy
17	Repulsion energy
18	Atomization energy

energy was set to zero. The energies of the other conformations of the same dimer were also corrected accordingly. This merely is the standard practice in most computational chemistry applications, to get away from total energies (that are irrelevant in most cases, and also frequently method-dependent). Here it focuses the machine-learning approach on the relevant relative energies between conformers.

By the above-mentioned treatment of the CCSD(T) conformational energies, the computed representations also were corrected the same way, i.e. the deviations of computed representations for each conformer with respect to the one with the lowest total energy were considered as the inputs to the machine-learning models.

For evaluation of solvation free energies, considering that the reference data were defined as the difference between the free energies of two liquid and gas states, we did not employ the same modification of energy and representations. However, in addition to the computed ImPerHam representations, we also considered the dielectric constant of solvents in each solute-solvent mixture as additional model input, to characterize the specific solvent under study, similar to conventional continuum solvation models and our recently developed machine-learning model [36].

To compare the accuracy of solvation free energies obtained by machine learning with those obtained via conventional solvation models, we also computed the solvation free energies via the widely-accepted SMD [50], PCM [51], and CPCM [52] continuum solvation models at the B3LYP/6-311+g(2d,p) level of theory in Gaussian 16 [53]. To that end, the geometry of the solutes was first optimized and then used for the computation of solvation free energy.

9.2.3 Developing machine-learning models

For more efficient and tractable training of machine-learning models, we employed variable selection as a commonly used approach in developing machine-learning models. To that end, the initial pool of computed representations was screened employing the MRMR variable selection algorithm [54] to yield different sets of input variables containing 4 to 45 representations.

After computation and screening of the required representations based on the above recipe, the machine-learning models taking those representations were set up to study the applications of interest.

Considering that the inhibition of CYP450 enzyme by potential drug candidates is a classification problem, i.e. the role of machine learning is only to classify the studied molecules as inhibitor or non-inhibitor, we employed the Support-Vector-Machine (SVM) classification [55] as a rigorous machine-learning method developed for this purpose. To that

end and to be able to make a reliable comparison to the large number of machine-learning models studied by Novotarskyi et al. [44] for the same purpose, we used the same training and test samples employed in that work.

For evaluation of conformational energy and solvation free energy, we employed artificial neural networks as a highly efficient machine-learning tool to map the dependency between required quantities and the proposed representations. To that end, training of the ANN models was carried out based on the guidelines provided in our recent study [56].

Accordingly, we assigned 60% of the dataset for training, 15% for validation, and 25% for testing the machine-learning models. We only studied small neural networks with one hidden layer and a very limited number of neurons. The upper limit of the selected number of neurons was assigned to allow having roughly 10 training samples or more per ANN constant. In many neural network models, much lower ratios of training samples to ANN constants are commonly employed, which can result in much higher accuracies for the obtained results. However, this may lead to reducing the extrapolation capability of the neural networks [56]. Therefore, we only considered a limited number of neurons with the upper limit discussed above.

For the machine-learning models developed to evaluate conformational energies, training, validation, and test sets were assigned based on the individual dimer types and not individual conformers, to avoid unreliable high-accuracy results due to interpolation. In other words, if a dimer was assigned to a training, validation, or test set, all its conformers were also assigned to the same set.

9.3 Results and discussion

9.3.1 Evaluation of CYP450 inhibition

By training SVM models via 3745 training samples provided by Novotarskyi et al. [44], the predictability of inhibitor activity of 3740 test set molecules was evaluated using the developed machine-learning models. According to the results, the best prediction of inhibitor activity for the training samples was observed for an SVM model that employed 16 representations as model inputs. Via this model, the inhibitor activity of test set compounds could be predicted with 79.6% accuracy which is very close to the results reported for the same datasets by the most accurate models studied by Novotarskyi et al. [44]. However, the models studied by Novotarskyi et al. employed a much greater number of representations, ranging from several hundred to roughly five thousand representations.

Considering that our employed representations are generated only for a few solvents, it is expected that for further extensions of the number of solvents, a more accurate evaluation of CYP450 inhibition can be achievable.

9.3.2 Machine-learning approximation of conformational energies via ImPerHam representations

By employing the GFN2-xTB semiempirical method to approximate relative energies of different conformers, we initially obtained an MUE of 22.164 kcal/mol compared to the CCSD(T) reference energies.

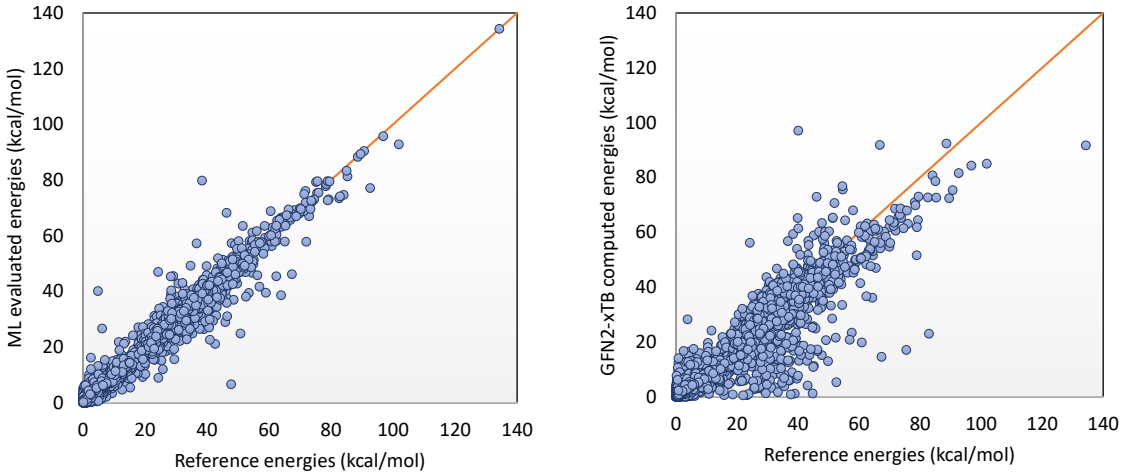


Fig. 9.2 Comparison of reference CCSD(T) energies for dimer conformers with energies evaluated by machine learning based on ImPerHam representations and computed by GFN2-xTB, respectively. Employing machine learning and ImPerHam representations improves the MUE from 22.164 kcal/mol to 1.400 kcal/mol, as a single-number measure of the overall deviation from the ideal line indicated in these two graphs.

Despite this large MUE of the originally evaluated energies by GFN2-xTB computations, a remarkable improvement in the accuracy of predicted energies was achieved via the developed machine-learning models based on different combinations of ImPerHam representations. According to the results, the studied machine-learning models could yield an MUE of 1.400 kcal/mol via a neural network model taking 38 ImPerHam representations as model inputs and 16 neurons in the hidden layer. These results show an improvement in originally computed energies based on the GFN2-xTB method by one order of magnitude. A comparison of the reference CCSD(T) configuration energies and the machine-learned

and GFN2-xTB computed ones are depicted in figure 9.2. As a further illustration, for the three dimers with the greatest energy variations between their conformers, the CCSD(T) energies are compared with machine learning and GFN2-xTB evaluated energies in different conformers in figure 9.4.

9.3.3 Machine-learning estimation of solvation free energies via ImPerHam representations

Among the studied machine-learning models for estimation the solvation free energies, the best result was obtained for a neural network with only 5 neurons in the hidden layer and 22 ImPerHam representations and solvent dielectric constants as model inputs. For this model, we obtained an MUE of 0.545 kcal/mol. Compared to the accuracy of the original ALPB, for which an MUE of 1.4 kcal/mol has been reported [47], our results show a substantial improvement by almost a factor of three. Further comparison of the obtained results with conventionally accepted continuum solvation models is reported in table 9.2. According to these results, the machine-learning evaluation of the solvation free energy via ImPerHam representations is significantly more accurate than SMD, PCM, and CPCM as the most extensively applied continuum solvation models, though obtained for a computational cost reduced by several orders of magnitude (a few seconds compared to several minutes to hours on a normal desktop PC). The distribution of solvation free energies predicted via the newly proposed machine-learning method and by the SMD solvation model, in comparison to experimentally determined data, is depicted in figure 9.3.

Among the other solvation models reported in table 9.2, only for the commercially available COSMO-RS solvation model higher accuracies have been reported. However, it should be noted that the results reported here are presented only as proof of concept. By trying more extensive sets of solvents, larger neural networks, and more demanding training of neural network models to search for the global minimum in the evaluated MUE, more accurate results by the machine-learning models are expected to be achievable.

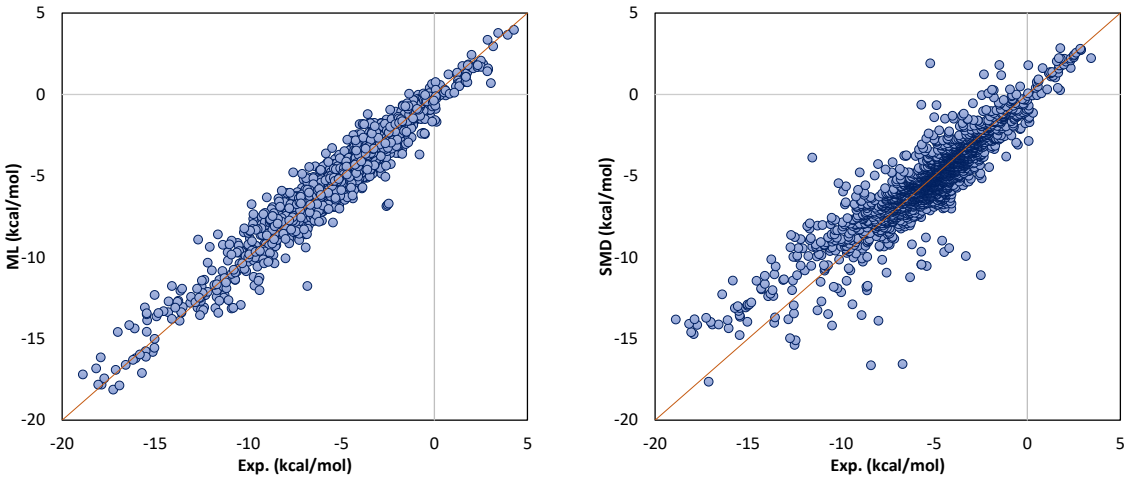


Fig. 9.3 Comparison of solvation free energies predicted via ML and SMD methods and experimental data.

Table 9.2 Comparison of the results of the new method with other models for solvation free energy prediction.

Method	Source	Nr. Samples	Nr. Solvents	Nr. Solutes	Deviation measure	Deviation (kcal/mol)
Machine learning	Present study	2493	91	435	MUE RMSE	0.545 0.739
COSMO-RS	Klamt-Diedenhofen [57]	2346	91	318	MUE RMSE	0.42145 0.69644
SM12	Marenich et al. [58]	2403	91	352	MUE	0.5457-0.6717
DCOSMO-RS	Klamt-Diedenhofen [57]	2346	91	318	MUE RMSE	0.6584 0.99724
Feature Functional Theory	Wang et al. [59]	668	1 (water)	668	RMSE	1.05
kernel-based ML	Rauer-Bereau [60]	355	1 (water)	355	MUE	1.06
atoms-in-molecules ANN	Zubatyuk et al. [61]	—	—	414	MUE	1.1
Structure-Property Relationship	Hutchinson-Kobayashi [62]	—	1 (water)	—	RMSE	1.65
SMD	Present study	2493	91	435	MUE RMSE	0.78623 1.1633
CPCM	Present study	2493	91	435	MUE RMSE	2.6942 3.1733
PCM	Present study	2493	91	435	MUE RMSE	2.9054 3.3948

9.3.4 Analysis of the performance of studied solvents and representations

As discussed in section 9.2.2, we considered the gas medium as well as 9 solvents, which provided a diverse range of dielectric constants and therefore different perturbations of the molecular Hamiltonians. The resulting energy attributes were considered as inputs for machine-learning models. The studied representations introduced in table 9.1 were also a very limited subset of potential energy attributes that could be defined and used as molecular representations. By analysis of those developed models that were among the top 10% of all studied models in terms of accuracy, we investigated the performance of individual solvents and representations.

The percentage of presence of each one of the studied mediums in the selected models with the highest accuracies is depicted in figure 9.5.

As can be seen in figure 9.5, the most effective perturbing mediums are not the same in different case studies. For example, while the energy attributes perturbed by benzene are available in 96% and 92% of the most effective machine-learning models developed for evaluating molecular energies and CYP450 inhibition, respectively, they are present in only 30% of the models which are effective in the estimation of solvation free energy. Additionally, the unperturbed Hamiltonian energy attributes as well as those obtained via perturbed Hamiltonian by ethyl acetate, water, and octane, are the most efficient ones in all considered case studies. The other interesting observation here is that all the studied solvents are employed in at least one-third of the models.

The percentage of presence of considered energy attributes listed in table 9.1 in the selected models is reported in table 9.3. Similar to the studied solvents, here also these results show the significant diversity in the efficiency of various energy attributes depending on the application. For example, while the representation with id equal to 2 is present in 87.5% of the accurate models developed for evaluation of molecular energies, it has not been employed in any of the selected models for evaluation of solvation free energy or inhibition of CYP450. Similarly, in all case studies, we can find solvents or representations which are present in all of the selected models. Therefore, studying a wider range of potential solvents or representations gives both high flexibility of ImPerHam representations in studying various problems and allows for achieving higher accuracies.

As can be seen in figure 9.5, accurate prediction of solvation free energy by machine-learning models can be achieved for a significantly lower number of solvents. The potential reason for that can be attributed to the fact that solvation free energy mainly depends on the dielectric constant of the solvent and the geometrical shape of the solute molecules [63]. While the former is already present in all machine-learning models as model input, as

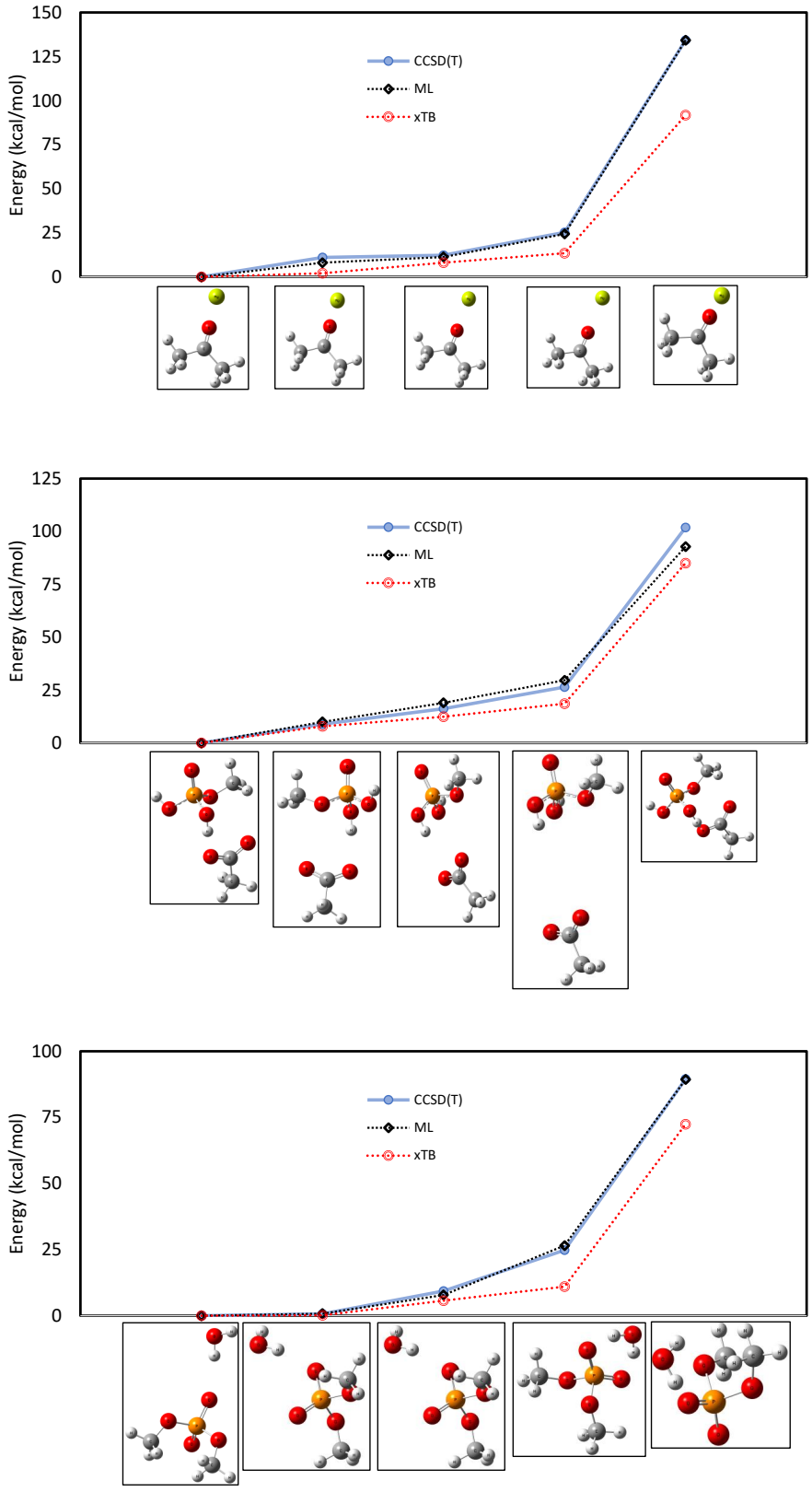


Fig. 9.4 Comparison of CCSD(T) energies with machine learning (ML) and GFN2-xTB evaluated energies for different conformers of three studied dimers with highest variation of energy between conformers. In the illustrated configurations, atoms of C, O, H, P, and Mg are presented with gray, red, white, orange, and green, respectively.

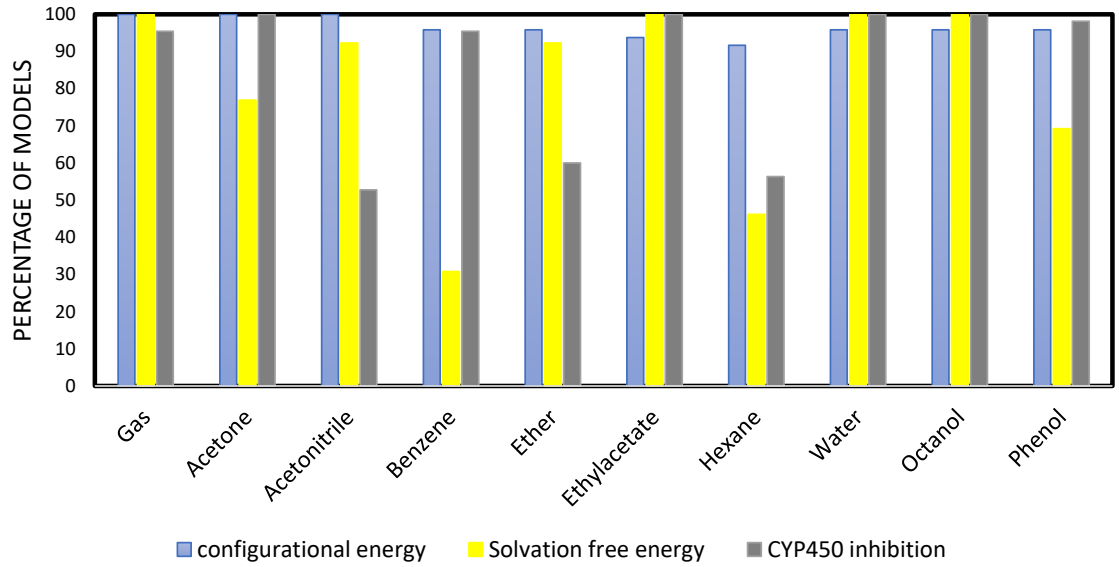


Fig. 9.5 Percentage of the presence of studied mediums in selected models for different considered applications.

discussed in the method section, the latter remains the same for different solvents. For the same reason, all of the machine-learning models that yield the best results for evaluation of solvation free energy employ total Gibbs free energy and cavitation Gibbs free energy as model inputs, as can be seen in table 9.3.

On the other hand, for the prediction of conformational energy, the potential energy attributes, such as isotropic electrostatic energy and the norm of the gradient vector, which is a measure of the energy difference between the molecular structure under study and the most stable conformation, are obviously the most relevant representations to the total conformational energy. Consequently, these energy attributes are present in 100% of the machine-learning models that yield the highest accuracies for the evaluation of conformational energies. Additionally, unlike the geometrical shapes required by the evaluation of solvation free energy, such energy attributes vary more significantly in different solvents. As a result, the highest employment of all studied solvents also is observed for this application.

For inhibition of the CYP450, as it is implied from the physics of the problem, the free energy of solvation of drug candidates in different solvents can determine the free energy of docking the drugs to the active site of the enzyme [64,65]. To that end, the most widely considered solvents are octanol and water, commonly employed as octanol-water partition coefficient [66-70]. Very interestingly, these two solvents are also present in all of the developed machine-learning models with the highest efficiency in evaluating the CYP450 inhibition. For the same reason, the energy attributes which are most clearly related to the

Table 9.3 Percentage of the presence of studied energy attributes in selected models in different applications.

Representation ID	Conformational energy	Solvation free energy	CYP450 inhibition
1	87.50	15.38	25.45
2	87.50	0.00	0.00
3	16.67	61.57	71.82
4	79.17	84.62	71.82
5	79.17	38.46	100
6	75.00	46.15	0.00
7	93.75	100	57.27
8	37.50	15.38	100
9	100	100	100
10	45.83	92.31	100
11	16.67	0.00	0.00
12	100	84.62	44.55
13	100	15.38	46.36
14	79.17	69.23	71.82
15	16.67	15.38	97.27
16	37.50	92.31	44.55
17	8.33	46.15	13.64
18	79.17	100	0.00

solvation free energy, such as electric, cavitation, and hydrogen bonding components of the solvation free energy, are present in 100% of the most effective machine-learning models developed for this specific application. From that perspective and considering that these energy attributes construct the total solvation free energy, the representation encoding the total solvation free energy becomes redundant and is not employed as extensively as its constituting components.

9.4 Conclusion

In the present study, we introduced ImPerHam as highly versatile representations for describing molecular systems and developing advanced machine-learning models. The ImPerHam representations are various energy attributes that are computed via molecular Hamiltonians in vacuum as well as in implicitly defined solvents. We demonstrated high efficiency and accuracy of machine-learning models based on ImPerHam representations in three diverse applications of predicting CYP450 inhibition by candidate molecules, evaluating the conformational energies in multi-atomic systems, and solvation free energies of diverse

solute-solvent mixtures. Our results have shown the capability of ImPerHam representations in developing transferable machine-learning models applicable to diverse molecular systems.

References:

1. Faulon, J.-L.; Faure, L., In silico, in vitro, and in vivo machine learning in synthetic biology and metabolic engineering. Current Opinion in Chemical Biology 2021, 65, 85-92.
2. Liu, J.; Li, J.; Wang, H.; Yan, J., Application of deep learning in genomics. Science China Life Sciences 2020, 1-19.
3. Lavecchia, A., Machine-learning approaches in drug discovery: methods and applications. Drug discovery today 2015, 20 (3), 318-331.
4. Sommer, C.; Gerlich, D. W., Machine learning in cell biology–teaching computers to recognize phenotypes. Journal of cell science 2013, 126 (24), 5529-5539.
5. Berka, K.; Srsen, S.; Slavicek, P., Is Machine Learning the Future of Theoretical Chemistry? CHEMICKE LISTY 2018, 112 (10), 640-647.
6. Liu, Y.; Yang, Q.; Li, Y.; Zhang, L.; Luo, S., Application of Machine Learning in Organic Chemistry. CHINESE JOURNAL OF ORGANIC CHEMISTRY 2020, 40 (11), 3812-3827.
7. Dral, P. O., Quantum chemistry in the age of machine learning. The journal of physical chemistry letters 2020, 11 (6), 2336-2347.
8. Schütt, K.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J., Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. Nature communications 2019, 10 (1), 1-10.
9. Gormley, A. J.; Webb, M. A., Machine learning in combinatorial polymer chemistry. Nature Reviews Materials 2021, 1-3.
10. Pflüger, P. M.; Glorius, F., Molecular machine learning: the future of synthetic chemistry? Angewandte Chemie International Edition 2020, 59 (43), 18860-18865.
11. Gamson, W.; Watson, K. In National Petroleum News 36, Tech. Sect, 1944; p 1944.
12. Joback, K. G.; Reid, R. C., Estimation of pure-component properties from group-contributions. Chemical Engineering Communications 1987, 57 (1-6), 233-243.
13. Fredenslund, A., Vapor-liquid equilibria using UNIFAC: a group-contribution method. Elsevier: 2012.
14. Alibakhshi, A.; Mirshahvalad, H.; Alibakhshi, S., A modified group contribution method for accurate prediction of flash points of pure organic compounds. Industrial & Engineering Chemistry Research 2015, 54 (44), 11230-11235.
15. He, T.; Li, S.; Chi, Y.; Zhang, H.-B.; Wang, Z.; Yang, B.; He, X.; You, X., An adaptive distance-based group contribution method for thermodynamic property prediction. Physical Chemistry Chemical Physics 2016, 18 (34), 23822-23830.
16. Alibakhshi, A.; Mirshahvalad, H.; Alibakhshi, S., Prediction of flash points of pure organic compounds: Evaluation of the DIPPR database. Process Safety and Environmental Protection 2017, 105, 127-133.

17. Li, R.; Herreros, J. M.; Tsolakis, A.; Yang, W., Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. Fuel 2020, 280, 118589.
18. Kibler, R.; Mohrhardt, B.; Zhang, M.; Breitner, L.; Howe, K. J.; Minakata, D., Group Contribution Method to Predict the Mass Transfer Coefficients of Organics through Various RO Membranes. Environmental science & technology 2020, 54 (8), 5167-5177.
19. Graziano, B.; Burkardt, P.; Neumann, M.; Pitsch, H.; Pischinger, S., Development of a Modified Joback–Reid Group Contribution Method to Predict the Sooting Tendency of Oxygenated Fuels. Energy & Fuels 2021.
20. Clark, J. A.; Santiso, E. E., SAFT-γ-Mie Cross-Interaction Parameters from Density Functional Theory-Predicted Multipoles of Molecular Fragments for Carbon Dioxide, Benzene, Alkanes, and Water. The Journal of Physical Chemistry B 2021, 125 (15), 3867-3882.
21. Fayaz-Torshizi, M.; Müller, E. A., Coarse-grained molecular dynamics study of the self-assembly of polyphilic bolaamphiphiles using the SAFT-γ Mie force field. Molecular Systems Design & Engineering 2021.
22. Lobanova, O.; Mejia, A.; Jackson, G.; Mueller, E. A., SAFT-γ force field for the simulation of molecular fluids 6: Binary and ternary mixtures comprising water, carbon dioxide, and n-alkanes. The Journal of Chemical Thermodynamics 2016, 93, 320-336.
23. Avendano, C.; Lafitte, T.; Galindo, A.; Adjiman, C. S.; Jackson, G.; Mueller, E. A., SAFT-γ force field for the simulation of molecular fluids. 1. A single-site coarse grained model of carbon dioxide. The Journal of Physical Chemistry B 2011, 115 (38), 11154-11169.
24. Behler, J., Perspective: Machine learning potentials for atomistic simulations. The Journal of chemical physics 2016, 145 (17), 170901.
25. Rupp, M., Machine learning for quantum mechanics in a nutshell. International Journal of Quantum Chemistry 2015, 115 (16), 1058-1073.
26. Behler, J.; Parrinello, M., Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical review letters 2007, 98 (14), 146401.
27. Bartok, A. P.; Payne, M. C.; Kondor, R.; Csányi, G., Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical review letters 2010, 104 (13), 136403.
28. Bartók, A. P.; Kondor, R.; Csányi, G., On representing chemical environments. Physical Review B 2013, 87 (18), 184115.
29. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A., Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters 2012, 108 (5), 058301.
30. Unruh, D.; Meidanshahi, R. V.; Goodnick, S. M.; Csányi, G.; Zimányi, G. T., Training a Machine-Learning Driven Gaussian Approximation Potential for Si-H Interactions. arXiv preprint arXiv:2106.02946 2021.
31. Liu, Y.-B.; Yang, J.-Y.; Xin, G.-M.; Liu, L.-H.; Csànyi, G.; Cao, B.-Y., Machine learning interatomic potential developed for molecular simulations on thermal properties of β-Ga2O3. The Journal of Chemical Physics 2020, 153 (14), 144501.
32. Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A., An accurate and transferable machine learning potential for carbon. The Journal of Chemical Physics 2020, 153 (3), 034702.
33. Davidson, E.; Daff, T.; Csanyi, G.; Finnis, M., Grand canonical approach to modeling hydrogen trapping at vacancies in α- Fe. Physical Review Materials 2020, 4 (6), 063804.
34. Behler, J., Constructing high-dimensional neural network potentials: A tutorial review. Interna-

tional Journal of Quantum Chemistry 2015, 115 (16), 1032-1050.
35. Behler, J., Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. Physical Chemistry Chemical Physics 2011, 13 (40), 17930-17955.
36. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. Nature Communications 2021,12(1),1-7.
37. Alibakhshi, A.; Hartke, B., Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. 2021.
38. Wu, Z.; Lei, T.; Shen, C.; Wang, Z.; Cao, D.; Hou, T., ADMET evaluation in drug discovery. 19. Reliable prediction of human cytochrome P450 inhibition using artificial intelligence approaches. Journal of chemical information and modeling 2019, 59 (11), 4587-4601.
39. Miners, J. O.; Mackenzie, P. I.; Knights, K. M., The prediction of drug-glucuronidation parameters in humans: UDP-glucuronosyltransferase enzyme-selective substrate and inhibitor probes for reaction phenotyping and in vitro–in vivo extrapolation of drug clearance and drug-drug interaction potential. Drug metabolism reviews 2010, 42 (1), 196-208.
40. Lasser, K. E.; Allen, P. D.; Woolhandler, S. J.; Himmelstein, D. U.; Wolfe, S. M.; Bor, D. H., Timing of new black box warnings and withdrawals for prescription medications. Jama 2002, 287 (17), 2215-2220.
41. Backman, J. T.; Wang, J. S.; Wen, X.; Kivistö, K. T.; Neuvonen, P. J., Mibefradil but not isradipine substantially elevates the plasma concentrations of the CYP3A4 substrate triazolam. Clinical Pharmacology & Therapeutics 1999, 66 (4), 401-407.
42. Porubek, D., CYP17A1: a biochemistry, chemistry, and clinical review. Current topics in medicinal chemistry 2013, 13 (12), 1364-1384.
43. Sansen, S.; Yano, J. K.; Reynald, R. L.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F., Adaptations for the Oxidation of Polycyclic Aromatic Hydrocarbons Exhibited by the Structure of Human P450 1A2. Journal of Biological Chemistry 2007, 282 (19), 14348-14355.
44. Novotarskyi, S.; Sushko, I.; Körner, R.; Pandey, A. K.; Tetko, I. V., A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. Journal of chemical information and modeling 2011, 51 (6), 1271-1280.
45. Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K., Quantum chemical benchmark databases of gold-standard dimer interaction energies. Scientific data 2021, 8 (1), 1-9.
46. Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G., Minnesota solvation database. Minnesota Solvation Database version 2012, 20.
47. Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S., A robust and efficient implicit solvation model for fast semiempirical methods. 2021.
48. Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S., Extended tight-binding quantum chemistry methods. Wiley Interdisciplinary Reviews: Computational Molecular Science 2021, 11 (2), e1493.
49. Bannwarth, C.; Ehlert, S.; Grimme, S., GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. Journal of chemical theory and computation 2019, 15 (3), 1652-1671.
50. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal solvation model based on solute electron

density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. The Journal of Physical Chemistry B 2009, 113 (18), 6378-6396.
51. Mennucci, B.; Cammi, R.; Tomasi, J., Excited states and solvatochromic shifts within a nonequilibrium solvation approach: A new formulation of the integral equation formalism method at the self-consistent field, configuration interaction, and multiconfiguration self-consistent field level. The Journal of chemical physics 1998, 109 (7), 2798-2807.
52. Barone, V.; Cossi, M., Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. The Journal of Physical Chemistry A 1998, 102 (11), 1995-2001.
53. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., Gaussian 16. Revision A 2016, 3.
54. Peng, H.; Long, F.; Ding, C., Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence 2005, 27 (8), 1226-1238.
55. Jakkula, V., Tutorial on support vector machine (svm). School of EECS, Washington State University 2006, 37.
56. Alibakshi, A., Strategies to develop robust neural network models: Prediction of flash point as a case study. Analytica chimica acta 2018, 1026, 69-76.
57. Klamt, A.; Diedenhofen, M., Calculation of solvation free energies with DCOSMO-RS. The Journal of Physical Chemistry A 2015, 119 (21), 5439-5445.
58. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Generalized born solvation model SM12. Journal of Chemical Theory and Computation 2013, 9 (1), 609-620.
59. Wang, B.; Wang, C.; Wu, K.; Wei, G. W., Breaking the polar-nonpolar division in solvation free energy prediction. Journal of computational chemistry 2018, 39 (4), 217-233.
60. Rauer, C.; Bereau, T., Hydration free energies from kernel-based machine learning: Compound-database bias. The Journal of Chemical Physics 2020, 153 (1), 014101.
61. Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O., Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. Science advances 2019, 5 (8), eaav6490.
62. Hutchinson, S. T.; Kobayashi, R., Solvent-specific featurization for predicting free energies of solvation through machine learning. Journal of chemical information and modeling 2019, 59 (4), 1338-1346.
63. Alibakhshi, A., Thermodynamically effective molecular surfaces for more efficient study of condensed-phase thermodynamics. 2021.
64. Azam, M. A.; Saha, N.; Jupudi, S., An explorative study on Staphylococcus aureus MurE inhibitor: induced fit docking, binding free energy calculation, and molecular dynamics. Journal of Receptors and Signal Transduction 2019, 39 (1), 45-54.
65. Oliveira, F. G.; Sant'Anna, C. M.; Caffarena, E. R.; Dardenne, L. E.; Barreiro, E. J., Molecular docking study and development of an empirical binding free energy model for phosphodiesterase 4 inhibitors. Bioorganic & medicinal chemistry 2006, 14 (17), 6001-6011.
66. Valencia-Islas, N. A.; Arguello, J. J.; Rojas, J. L., Antioxidant and photoprotective metabolites of Bunodophoron melanocarpum, a lichen from the Andean páramo. Pharmaceutical Sciences 2020, 27 (2), 281-290.
67. Ghamri, M.; Harkati, D.; Belaidi, S.; Boudergua, S.; Said, R. B.; Linguerri, R.; Chambaud, G.; Hochlaf, M., Carbazole derivatives containing chalcone analogues targeting topoisomerase II inhibi-

tion: First principles characterization and QSAR modelling. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2020, 242, 118724.
68. Stępnik, K.; Kukula-Koch, W., In Silico Studies on Triterpenoid Saponins Permeation through the Blood–Brain Barrier Combined with Postmortem Research on the Brain Tissues of Mice Affected by Astragaloside IV Administration. International journal of molecular sciences 2020, 21 (7), 2534.
69. Ventura, F. F.; Mendes, L. F.; Oliveira, A. G.; Bazito, R. C.; Bechara, E. J.; Freire, R. S.; Stevani, C. V., Evaluation of phenolic compound toxicity using a bioluminescent assay with the fungus Gerronema viridilucens. Environmental toxicology and chemistry 2020, 39 (8), 1558-1565.
70. Gunesch, A. P.; Zapatero-Belinchón, F. J.; Pinkert, L.; Steinmann, E.; Manns, M. P.; Schneider, G.; Pietschmann, T.; Brönstrup, M.; von Hahn, T., Filovirus antiviral activity of cationic amphiphilic drugs is associated with lipophilicity and ability to induce phospholipidosis. Antimicrobial agents and chemotherapy 2020, 64 (8), e00143-20.

Chapter 10

Summary and Outlook

In this thesis, we studied theoretical evaluation of thermochemistry and developed advanced methods which allow going beyond the state-of-the-art, in terms of accuracy or computational efficiency.
In this chapter, I summarize the major achievements acquired in these studies as well as my suggestions for potential improvements in future research.

10.1 Summary of the carried out projects

In chapter three, we employed static ab-initio calculations and normal mode analysis as the classical approaches for evaluating the enthalpy of combustion reactions. In addition to providing an accurate benchmark dataset for studying the intricacies of those classical approaches, one motivation for considering this case study was a surprising observation of remarkable inconsistency between theoretically predicted combustion enthalpies and experimentally measured data, reported in several studies. For enhanced theoretical computation of combustion enthalpy, we proposed and implemented corrections of non-ideality and phase change thermodynamics which were not previously considered elsewhere. With those considerations and a combination of thermodynamics and theoretical chemistry, we could achieve the highest accuracy ever reported for theoretical estimation of combustion enthalpy. We demonstrated that searching for the global minimum in the configurational energy of molecules as well as the accuracy of the employed level of theory in the evaluation of ground-state total electronic energies and not the thermal energies play the most important role in the accuracy of predicted enthalpies for gas-phase reactions. Based on the latter observation and as a suggestion for future research, we recommend investigating the performance of hybrid computations in which the ground-state electronic energies are evaluated via low-cost but accurate methods specifically developed for this purpose, for example, the

DLPNO method, and computing the thermal energies by other conventional methods, as a cost-effective approach.

In chapter four, we theoretically estimated the equilibrium constant of the boron isotope exchange reaction between boric acid and borate in pure water and seawater, as one of the quantities with significant geochemical importance. To that end, we employed three different approaches, namely static ab-initio computations with the explicit solvent approach, the same computations based on the implicit solvent approach, and thermodynamic perturbation of kinetic energy, calculated via the virial estimator within path-integral molecular dynamics simulations.

For the explicit solvent approach, we proposed partial normal mode analysis with frozen solvents as a highly effective and reliable method for both significantly reducing the computational cost and enhancing the reliability of the obtained results.

All these three diverse methods yielded consistent results in excellent agreement with the experimentally determined data, which confirmed the reliability of the newly developed methods.

The estimated equilibrium constant for the studied boron isotope exchange reaction is highly required by empirical models that reconstruct ancient seawater pH and atmospheric pCO2 by analyzing boron isotopic ratios in fossils of marine shells. As a suggestion for future studies, I recommend employing the developed methods for a direct evaluation of the rate and equilibrium constant of the incorporation of boron species and the fractionation of boron isotopes in the carbonate shells, i.e. directly simulating the solid-liquid equilibrium by theoretical studies instead of the indirect application of results estimated for the liquid-liquid equilibrium.

In chapter five, we developed a physical model for describing the temperature dependence of vaporization enthalpy and its connection to molecular surfaces, employing classical and statistical thermodynamics. We proposed an analytical relationship which is currently the only theoretically proposed model capable of high accuracy evaluating vaporization enthalpies of compounds from diverse chemical families and for a wide temperature range from melting points to the critical temperature. We extensively examined and demonstrated the veracity and accuracy of our proposed approach not only in the prediction of vaporization enthalpy but also for the evaluation of other thermodynamic quantities. The other major and controversial outcome of the proposed method was theoretically invalidating the other models, empirically proposed to relate solution thermodynamics and molecular surfaces, with a history of almost one century. We demonstrated the substantially higher accuracy and reliability of our theoretically proposed model compared to the previously employed ones.

As the other main achievement of that study, we theoretically defined and proposed the

thermodynamically effective molecular surfaces. We demonstrated that the theoretically defined thermodynamically effective surfaces were only slightly different from empirically proposed van-der-Waals surfaces. However, this slight deviation resulted in a substantial improvement in the predictability of multiple thermodynamic quantities. As a result, we proposed the thermodynamically effective surfaces as a reliable and more efficient alternative for the currently defined molecular surfaces, especially for studying condensed-phase thermodynamics.
In that work, we mainly focused on demonstrating the efficiency of those surfaces and potential applications. For their practical usage, I suggest parametrizing the currently defined computer algorithms for in-silico prediction of those surfaces as well as studying their performance for other applications, most specifically in life-science studies.

In the first part of this thesis including the scientific works discussed above, the main aim was method development employing the classic theoretical approaches and tools for enhanced study of thermochemistry. For the second part of this thesis, which included the methods introduced in the following, the main focus was to exploit the powerful capability of machine learning tools and to demonstrate their efficiency in studying very challenging problems in thermochemistry.

To that end, we first provided necessary guidelines for developing more reliable and accurate artificial neural network models as the main machine-learning tool employed in this thesis, in chapter six. We showed the importance of appropriately selecting the neural network details such as the number of hidden-layer neurons, training algorithms, initialization, transfer function, and validation strategies, which are commonly overlooked in many neural network models.
We benchmarked the provided guidelines via in-silico prediction of the flash-point of pure hydrocarbons based on the group contribution method, which yielded the most accurate results for that specific application, compared to results reported elsewhere by that time.

In chapter seven, inspired by the remarkable improvement in predictability of vaporization enthalpy obtained through the thermodynamically effective surfaces that we conceptualized in chapter five, we developed machine learning and analytical models for practical and high-accuracy prediction of vaporization enthalpy. We provided straightforward correlations which employ approximation of thermodynamically effective surfaces and provide accuracies comparable to the most accurate empirically developed correlations. Via the developed machine learning models, we could achieve the most accurate prediction of vaporization enthalpy ever reported for a large dataset of compounds and a large temperature range.

In chapter eight, we introduced the Machine learning Polarizable Continuum solvation Model (ML-PCM) for evaluation of solvation free energy, as one of the most important

thermophysical properties with extensive applications in a broad range of scientific disciplines, and also one of the general objectives of this thesis. Applying our new ML-PCM approach for a large dataset, we obtained one of the most accurate results ever reported for solvation free energy estimation, superseding the accuracy of the other most widely accepted solvation models, in some cases by almost one order of magnitude and for almost no additional computational cost. Our main suggestions for future improvements of these results are those already employed by the method presented in chapter nine of this thesis.

In chapter nine, we proposed ImPerHam representations as a novel class of highly versatile representations. We employed ImPerHam representations for developing machine learning models capable of cost-effective and accurate evaluation of inhibition of a CYP450 enzyme, solvation free energies, and configurational energy of molecules. For the inhibition of the CYP450 enzyme, we achieved accuracies that are comparable to the most accurate reported results for this purpose but employing a much smaller number of representations compared to those works. For the prediction of solvation free energies by the developed machine learning models, we obtained remarkably more accurate results for a significantly smaller computational cost, compared to the conventional continuum solvation models. For the configurational energy, using the new representations and machine learning, we could improve the accuracy of the GFN2-xTB semiempirical method by one order of magnitude and achieve reasonable accuracies compared to CCSD(T) reference data for a very large dataset. These three diverse applications, which are among the most challenging problems in the field, offer ImPerHam as a highly efficient method for generating versatile representations for different applications in molecular sciences.

10.2 Communicating science

To make our findings and achievements publicly available, the scientific works presented in chapters three to nine were considered for publication in peer-reviewed journals. Among them, two works have already been published by the date of submitting this thesis, and the other five are either currently under review in peer-reviewed journals or are in the process of being submitted shortly. Additionally, considering that implementation of the machine learning models usually requires profound technical knowledge on that specific topic, we provided C++ codes with detailed user instructions to allow a straightforward and convenient application by a broader range of readers for all of the presented machine learning models.

10.3 Outlook

In summary, our results demonstrated that both the classical methods and machine learning models can be employed as highly effective, reliable, and accurate approaches for most challenging applications in thermochemistry. The advantage of the former approach is that for a high enough accuracy of the employed level of theory and careful attention to their important intricacies, a highly accurate evaluation of thermochemistry is in most cases achievable. More importantly, the classical methods do not require several examples of the problem of interest which is required to train the machine learning models. It makes this approach the method of choice for tackling newly defined scientific challenges as well as problems for which extensive reference data are not readily available. The central reason for these advantages is that those methods deal with the exact physical rules governing the problems of interest. Nevertheless, the computational costs and challenges of employing classical tools are typically demanding and require profound knowledge in those fields, which can make them inaccessible for a broad range of scientists in other fields. It has indeed been the motivation of developing machine learning models to address these latter limitations. From that perspective, machine learning can be regarded as a unique tool possessing a number of advantages that are typically not achievable by classical approaches. As the most significant one, although developing reliably trained machine learning models can be a highly demanding and laborious procedure, once such models are developed, they can provide the same or better accuracies compared to classical approaches but for a computational cost that can be several orders of magnitude smaller. Additionally, unlike versatile classical approaches, implementing the machine learning models is becoming more and more convenient, even for non-expert users, with the help of a large number of currently available or forthcoming user-friendly software tools.

Nevertheless, as it was extensively elaborated in chapters six and nine, careful attention still should be taken to assure the reliability of the developed machine learning models. Although cross-validation has been extensively used for this purpose, as it was demonstrated in chapter six, this cannot always guarantee the reliability of the developed machine learning models. Therefore, developing and employing more rigorous validation strategies as well as efficient computer algorithms that can automate this procedure can be regarded as the most ambitious scientific goals and direction of future progress in that field, to the author's opinion.

Declaration

I, Amin Alibakhshi, hereby declare that the work presented in this thesis was done by me, under the supervision of Prof. Dr. Bernd Hartke, with no other help than the referenced sources in the text. This is my first dissertation and the work has never been used in any other dissertation attempts. I have never been deprived of an academic title. The dissertation complies to the good scientific practice rules as proposed by the German Research Foundation (DFG).

Amin Alibakhshi
September 2021