

# **Peptide-Cation Systems: Conformational Search, Benchmark Evaluation, and Force Field Parameter Adjustment Using Regularized Linear Regression**

THÈSE N° 8812 (2018)

PRÉSENTÉE LE 28 SEPTEMBRE 2018  
À LA FACULTÉ DES SCIENCES DE BASE  
INSTITUT DE PHYSIQUE  
PROGRAMME DOCTORAL EN PHYSIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

**Markus SCHNEIDER**

acceptée sur proposition du jury:

Prof. R. Houdré, président du jury  
Dr C. Baldauf, T. Rizzo, directeurs de thèse  
Prof. G. von Helden, rapporteur  
Dr L. Rulíšek, rapporteur  
Prof. M. Ceriotti, rapporteur



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

Suisse  
2018



# Abstract

Metal cations often play an important role in shaping the three-dimensional structure of peptides. As an example, the model system AcPheAla<sub>5</sub>LysH<sup>+</sup> is investigated in order to fully understand the forces that stabilize its helical structure. In particular, the question of whether the local fixation of the positive charge at the peptide's C-terminus is a prerequisite for forming helices is addressed by replacing the protonated lysine residue by alanine and a sodium cation. The combination of gas-phase cold-ion vibrational spectroscopy with molecular simulations based on density-functional theory (DFT) revealed that the charge localization at the C-terminus is imperative for helix formation in the gas phase as this stabilizes the structure through a cation-helix dipole interaction. For sodiated AcPheAla<sub>6</sub>, globular rather than helical structures were found caused by the strong cation-backbone and cation- $\pi$  interactions. Interestingly, the global minimum-energy structure from simulation is not present in the experiment where the system remains kinetically trapped in a solution-state structure.

Thereby calculated energies and IR spectra that are sufficiently accurate relied on DFT with computationally costly hybrid functionals, while for the structure search low-computational-cost force field (FF) models are crucial. This inspired a study where the goodness of commonly applied levels of theory, i.e. FFs, semi-empirical methods, density-functional approximations, composite methods, and wavefunction-based methods are being evaluated with respect to benchmark-grade coupled-cluster calculations. Acetylhistidine – either bare or in presence of a zinc cation – thereby serves as a molecular benchmark system. Neither FFs nor semi-empirical methods are reliable enough for a description of these systems within “chemical accuracy” of 1 kcal/mol. Accurate energetic description within chemical accuracy is achieved for all systems using the meta-GGA SCAN or computationally more demanding hybrid functionals. The double-hybrid functional B3LYP+XYG3 is best resembling the benchmark method DLPNO-CCSD(T).

Despite poor energetic performances of conventional FFs for peptides in the gas phase, their low computational costs still render them appealing tools for large-scale structure searches. Consequently, a machine learning approach is presented where the torsional parameters and (if desired) van der Waals parameters in the potential-energy function of a particular FF are adjusted by fitting it against DFT energies using regularized regression models like LASSO or Ridge regression. For the peptide AcAla<sub>2</sub>NMe, this resulted in a significant improvement when comparing to standard OPLS-AA FF parameters. For more challenging peptide-cation systems, e.g. AcAla<sub>2</sub>NMe + Na<sup>+</sup>, this approach does not give satisfying results, which is caused

---

by the formulation of the potential energy of the FF itself: While derived empirical partial charges using Hirshfeld partitioning or the electrostatic potential (ESP) decrease the accuracy, part of the energetic discrepancy can be “compensated” due to the flexibility of the torsional contributions in terms of the energetic description.

**Keywords:** peptide-cation systems, helical peptides, conformer-selective IR-UV spectroscopy, benchmark calculations, DFT, coupled-cluster, force fields, machine learning, Ridge regression, LASSO

# Zusammenfassung

Metallkationen spielen oft eine wichtige Rolle beim Formen dreidimensionaler Strukturen von Peptiden. Als Beispiel dafür wird das System  $\text{AcPheAla}_5\text{LysH}^+$  untersucht um die für die Stabilisierung helikaler Strukturen ursächlichen Kräfte zu verstehen. Im Detail wird der Frage nachgegangen, ob die Fixierung der lokalen positiven Ladung am C-Terminus des Peptids eine Voraussetzung für die Bildung der Helix ist, indem das protonierte Lysin-Residuum durch ein Alanin und ein Natrium-Kation ersetzt wird. Durch die Kombination von Kalte-Ionen-Vibrationspektroskopie im Vakuum und molekularen Simulationen basierend auf der Dichtefunktionaltheorie (DFT) wurde gezeigt, dass die lokale Ladung am C-Terminus zwingende Voraussetzung für die Helix-Bildung im Vakuum ist. Für das System  $\text{AcPheAla}_6 + \text{Na}^+$  wurden hingegen globuläre Strukturen gefunden, welche durch starke Kation-Rückgrat- und Kation- $\pi$ -Wechselwirkungen verursacht werden. Die in der Simulation gefundene Struktur globaler minimaler Energie wurde im Experiment nicht beobachtet, weil das System in einer Lösungs-Struktur kinetisch gefangen bleibt.

Für ausreichend genau berechnete Energien und IR-Spektren benötigt man dabei rechenaufwändige DFT-Hybridfunktionale, während für die Struktursuche Kraftfeld-Modelle geringem Rechenaufwands verwendet werden. Dieser Umstand motivierte eine Benchmark-Studie, in der die Qualität gängiger theoretischer Methoden, d.h. Kraftfelder, semi-empirische Methoden, Dichtefunktionalnäherungen, Mischmethoden und Methoden basierend auf Wellenfunktionen, gegen Coupled-Cluster-Rechnungen getestet werden. Acetylhistidin, mit und ohne einem angrenzenden Zink-Kation, dient dabei als molekulares Benchmark-System. Weder Kraftfelder noch semi-empirische Methoden sind dabei verlässlich genug solche Systeme innerhalb der „chemischen Genauigkeit“ von 1 kcal/mol zu beschreiben. Eine Beschreibung der Energie innerhalb der chemischen Genauigkeit wird für alle System bei Verwendung des meta-GGA SCAN- oder der rechenaufwändigeren Hybridfunktionale gefunden. Das Doppelhybridfunktional B3LYP+XYG3 beschreibt die Benchmark-Methode DLPNO-CCSD(T) am besten.

Trotz der ungenauen energetischen Beschreibung konventioneller Kraftfelder für Peptide im Vakuum, kommen diese wegen ihres niedrigen Rechenaufwands oft bei großangelegten Struktursuchen zum Einsatz. Diese Tatsache motivierte ein Machine-Learning-Verfahren, in dem Torsionsparameter und (falls gewünscht) van-der-Waals-Parameter in der Funktion der potenziellen Energie eines bestimmten Kraftfelds gegen DFT-Energien durch Einsatz regularisierter Regressionsmodelle wie Ridge-Regression oder LASSO gefittet werden. Für das Peptid  $\text{AcAla}_2\text{NMe}$  resultierte dies in einer signifikanten Verbesserung verglichen mit

---

den Standardwerten des OPLS-AA Kraftfeldes. Für kompliziertere Peptid-Kation-Systeme wie AcAla<sub>2</sub>NMe + Na<sup>+</sup> liefert das Verfahren keine zufriedenstellende Ergebnisse, wofür die Formulierung der potenziellen Energie des Kraftfelds selbst ursächlich ist: Während empirisch abgeleitete Partiaalladungen, entweder durch Anwendung der Hirshfeld-Partitionierung oder des elektrostatischen Potentials (ESP), zu ungenaueren Ergebnissen führen, kann ein Teil der energetischen Diskrepanz durch die Flexibilität der Torsionsterme in der energetischen Beschreibung „kompensiert“ werden.

**Schlagwörter:** Peptid-Kation-Systeme, Helikale Peptide, Konformer-selektive IR-UV Spektroskopie, Benchmark-Rechnungen, DFT, Coupled-Cluster, Kraftfelder, Machine Learning, Ridge-Regression, LASSO

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Zusammenfassung</b>	<b>v</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Tables</b>	<b>xi</b>
<b>List of Acronyms</b>	<b>xiv</b>
<b>1 Motivation and Overview</b>	<b>1</b>
<b>2 Theoretical and Experimental Background, Methods, and Techniques</b>	<b>5</b>
2.1 Amino Acids, Peptides, and Proteins . . . . .	6
2.2 The Structure and Energetics of Peptides in the Gas Phase . . . . .	11
2.3 Description of the Potential Energy Surface (PES) . . . . .	20
2.3.1 Force Fields . . . . .	20
2.3.2 Schrödinger Equation and Born-Oppenheimer Approximation . . . . .	24
2.3.3 Variational Principle for the Ground State and Hartree-Fock Method . . . . .	26
2.3.4 Semi-Empirical Quantum Chemistry Methods . . . . .	30
2.3.5 Density-Functional Theory . . . . .	33
2.3.6 <i>A posteriori</i> van der Waals Correction Schemes in Density-Functional Theory and Semi-Empirical Quantum Chemistry Methods . . . . .	42
2.3.7 Electron Correlation and Møller-Plesset Perturbation Theory . . . . .	46
2.3.8 Configuration Interaction and Coupled-Cluster Theory . . . . .	49
2.3.9 Computer Simulations and Practical Considerations . . . . .	53
2.4 Conformational Sampling and Basin-Hopping . . . . .	55
2.5 Description of the Free Energy Surface and Comparison to Experiment . . . . .	58
2.5.1 Infrared Spectra and Free Energy Calculations in Harmonic Approximation	58
2.5.2 Experimental Setup . . . . .	67
<b>3 Conformational Structure Search and Kinetically Trapped Liquid-State Conformers of a Sodiated Model Peptide Observed in the Gas Phase</b>	<b>69</b>
3.1 Motivation: Prerequisites of Helix Formation in the Gas Phase . . . . .	70
3.2 Experimental Setup . . . . .	72

## Contents

---

3.3	Computational Methods . . . . .	72
3.4	Results and Discussion . . . . .	73
3.4.1	AcPheAla <sub>5</sub> LysH <sup>+</sup> . . . . .	73
3.4.2	AcPheAla <sub>6</sub> + Na <sup>+</sup> . . . . .	77
3.5	Conclusion . . . . .	82
<b>4</b>	<b>Energetics and Benchmark of Across-the-scale Energy Methods of Acetyl-Histidine Protomers with and without Zn<sup>2+</sup></b>	<b>83</b>
4.1	Motivation and Overview . . . . .	84
4.2	Computational Details . . . . .	86
4.2.1	Conformational Sampling . . . . .	86
4.2.2	Levels of Theory and Energy Calculation Methods . . . . .	87
4.2.3	Mean Absolute Error (MAE) and Maximum Error (ME) . . . . .	91
4.3	Results . . . . .	92
4.3.1	Energy Hierarchies . . . . .	92
4.3.2	Selection of Minima Structures . . . . .	96
4.3.3	Validation of DLPNO-CCSD(T) as the Reference Method . . . . .	96
4.3.4	Benchmarking Force Fields and Semi-Empirical Methods . . . . .	97
4.3.5	Benchmarking Standard DFAs and Methods Beyond . . . . .	100
4.3.6	Considering Calculation Times . . . . .	102
4.4	Conclusions . . . . .	103
<b>5</b>	<b>Force Field Parameterization Using Regularized Linear Regression</b>	<b>105</b>
5.1	Motivation . . . . .	106
5.2	Computational Details and Framework . . . . .	108
5.2.1	Functional Form and Parameters of Empirical Force Fields . . . . .	108
5.2.2	Regularized Linear Regression: Ridge Regression and LASSO . . . . .	110
5.2.3	The “Framework For Adjusting Force Fields Using Regularized Regression” (FFAFFURR) . . . . .	114
5.3	Results . . . . .	121
5.4	Conclusion and Outlook . . . . .	135
<b>6</b>	<b>Summary</b>	<b>139</b>
<b>A</b>	<b>Appendix: Listing of Force Field Parameters for AcAla<sub>2</sub>NMe + Na<sup>+</sup></b>	<b>143</b>
	<b>Bibliography</b>	<b>149</b>
	<b>Acknowledgments</b>	<b>179</b>
	<b>Curriculum Vitae</b>	<b>181</b>
	<b>Publications and Conference Contributions</b>	<b>182</b>



# List of Figures

1.1	Two examples of peptide-cation interaction sites. . . . .	2
2.1	Structural formulas of the general form of the 20 DNA encoded amino acids and their stereochemistry. . . . .	6
2.2	Ionic and tautomeric forms of the histidine side chain at physiologically relevant pH. . . . .	9
2.3	Depiction of the condensation of two amino acids to form a peptide bond. . . . .	9
2.4	Schematic representation of the exemplary zwitterionic peptide Ala <sub>5</sub> . . . . .	10
2.5	Examples of primary, secondary, and tertiary structure of peptides and proteins. . . . .	15
2.6	Schematic view of hydrogen bond patterns in different helix types. . . . .	16
2.7	Schematic view of hydrogen bond patterns in $\beta$ -sheets. . . . .	16
2.8	Schematic depiction of a funnel-shaped free energy landscape. . . . .	18
2.9	Schematic illustration of the basin-hopping approach. . . . .	56
2.10	Schematic illustration of the cold ion spectroscopy instrument. . . . .	68
3.1	Illustration of helix-stabilizing factors for peptides in the gas phase. . . . .	71
3.2	Structural formulas of AcPheAla <sub>5</sub> LysH <sup>+</sup> and AcPheAla <sub>6</sub> + Na <sup>+</sup> . . . . .	71
3.3	Comparison of energy hierarchies of conformers of AcPheAla <sub>5</sub> LysH <sup>+</sup> between the conformational search applied here and the search performed by Rossi <i>et al.</i> . . . . .	74
3.4	Energy hierarchies of conformers of AcPheAla <sub>5</sub> LysH <sup>+</sup> . . . . .	75
3.5	Relative DFT energies and Helmholtz free energy hierarchies as well as structural illustrations and corresponding measured and calculated IR spectra for the lowest-energy conformers of AcPheAla <sub>5</sub> LysH <sup>+</sup> . . . . .	76
3.6	Measured UV spectrum for the system of AcPheAla <sub>5</sub> LysH <sup>+</sup> . . . . .	77
3.7	Energy hierarchies of conformers of AcPheAla <sub>5</sub> LysH <sup>+</sup> . . . . .	78
3.8	Relative DFT energies and Helmholtz free energy hierarchies as well as structural illustrations and corresponding measured and calculated IR spectra for the lowest-energy conformers of AcPheAla <sub>6</sub> + Na <sup>+</sup> . . . . .	79
3.9	Measured UV spectrum for the system of AcPheAla <sub>6</sub> + Na <sup>+</sup> . . . . .	80
3.10	Comparison of energy hierarchies on the PES between gas-phase calculations and calculations including implicit solvation effects. . . . .	81
3.11	For the system of AcPheAla <sub>6</sub> + Na <sup>+</sup> , the two measured conformer-selective IR spectra with lowest intensity are compared to vibrational calculations. . . . .	82

## List of Figures

---

4.1	Chemical structures of AcH showing the possible protonation states. . . . .	85
4.2	Example of a correlation plot of two different sets of conformers. . . . .	92
4.3	Obtained energy hierarchies for negatively charged and neutral AcH, bare and with an additional $Zn^{2+}$ . . . . .	93
4.4	Illustration of the structure of the lowest-energy conformer for each depicted protonation state. . . . .	94
4.5	Correlation plots for benchmarking DLPNO-CCSD(T) against conventional CCSD(T). . . . .	98
4.6	Mean absolute errors and maximum errors for different force fields and semi-empirical methods with respect to DLPNO-CCSD(T). . . . .	99
4.7	Mean absolute errors and maximum errors for different standard DFAs, the composite method PBEh-3c, double hybrid DFA B3LYP+XYG3, and the wavefunction-based MP2 method with respect to DLPNO-CCSD(T). . . . .	101
5.1	Schematic illustration of the selected volume used for evaluating the electrostatic potential (ESP). . . . .	117
5.2	Structural formula and illustration of AcAla <sub>2</sub> NMe. . . . .	121
5.3	Illustration of the standard TINKER atom types of the OPLS-AA force field for the system of AcAla <sub>2</sub> NMe. . . . .	123
5.4	MAEs and MEs for the test set of AcAla <sub>2</sub> NMe when multiplying the obtained Hirshfeld and ESP partial charges with scaling factors. . . . .	125
5.5	Distributions of Hirshfeld and ESP charges over the training set of conformers for the system of AcAla <sub>2</sub> NMe. . . . .	126
5.6	Estimated regression coefficients $\epsilon_{ij}$ , the RSS, and calculated MAEs and MEs for the test set of AcAla <sub>2</sub> NMe obtained by using LASSO regression. . . . .	129
5.7	Estimated regression coefficients $V_1^{ij}$ , $V_2^{ij}$ , and $V_3^{ij}$ , the RSS, and calculated MAEs and MEs for the test set of AcAla <sub>2</sub> NMe obtained by using Ridge regression. . . . .	130
5.8	Estimated regression coefficients $V_1^{ij}$ , $V_2^{ij}$ , and $V_3^{ij}$ , the RSS, and calculated MAEs and MEs for the test set of AcAla <sub>2</sub> NMe obtained by using LASSO regression. . . . .	132
5.9	Exemplary illustration of conformers of AcAla <sub>2</sub> NMe + Na <sup>+</sup> for which the sodium cation is surrounded by a varying number of oxygen atoms. . . . .	134
A.1	Distributions of Hirshfeld and ESP charges over the training set of conformers for the system of AcAla <sub>2</sub> NMe + Na <sup>+</sup> . . . . .	145
A.2	Estimated regression coefficients $\epsilon_{ij}$ , the RSS, and calculated MAEs and MEs for the test set of AcAla <sub>2</sub> NMe + Na <sup>+</sup> obtained by using LASSO regression. . . . .	146
A.3	Estimated regression coefficients $V_1^{ij}$ , $V_2^{ij}$ , and $V_3^{ij}$ , the RSS, and calculated MAEs and MEs for the test set of AcAla <sub>2</sub> NMe + Na <sup>+</sup> obtained by using LASSO regression. . . . .	147

# List of Tables

2.1	Summary of the 20 natural amino acids. . . . .	7
4.1	Minima selection criteria across the tackled systems and protonation states. . .	96
5.1	Summary of chemical symbols, atom types, and classes for the system of AcAla <sub>2</sub> NMe using the standard TINKER notation of the OPLS-AA force field. . .	122
5.2	Original OPLS-AA FF parameters $r_{ij}^0$ and $\theta_{ij}^0$ and their adjusted counterparts for the system of AcAla <sub>2</sub> NMe. . . . .	124
5.3	Original OPLS-AA FF parameters $q_i$ and their adjusted counterparts for the system of AcAla <sub>2</sub> NMe. . . . .	124
5.4	Original OPLS-AA FF parameters $\sigma_{ij}$ and their adjusted counterparts for the system of AcAla <sub>2</sub> NMe. . . . .	128
5.5	Original OPLS-AA FF parameters $\epsilon_{ij}$ and their adjusted counterparts for the system of AcAla <sub>2</sub> NMe using multiple linear regression. . . . .	128
5.6	Original OPLS-AA FF parameters $\epsilon_{ij}$ and their adjusted counterparts for the system of AcAla <sub>2</sub> NMe using LASSO regression. . . . .	129
5.7	Original OPLS-AA FF parameters $V_1^{ij}$ , $V_2^{ij}$ , and $V_3^{ij}$ as well as their adjusted counterparts for the system of AcAla <sub>2</sub> NMe using linear regression. . . . .	130
5.8	Original OPLS-AA FF parameters $V_1^{ij}$ , $V_2^{ij}$ , and $V_3^{ij}$ as well as their adjusted counterparts for the system of AcAla <sub>2</sub> NMe using linear regression. . . . .	131
5.9	Original OPLS-AA FF parameters $V_1^{ij}$ , $V_2^{ij}$ , and $V_3^{ij}$ as well as their adjusted counterparts for the system of AcAla <sub>2</sub> NMe using LASSO regression. . . . .	132
5.10	Overview on calculated MAEs and MEs for AcAla <sub>2</sub> NMe and AcAla <sub>2</sub> NMe + Na <sup>+</sup> . . .	135
A.1	Original OPLS-AA FF parameters $r_{ij}^0$ and $\theta_{ij}^0$ and their “adjusted” counterparts for the system of AcAla <sub>2</sub> NMe + Na <sup>+</sup> . . . . .	143
A.2	Original OPLS-AA FF parameters $q_i$ and their “adjusted” counterparts for the system of AcAla <sub>2</sub> NMe + Na <sup>+</sup> . . . . .	144
A.3	Original OPLS-AA FF parameters $\sigma_{ij}$ and their “adjusted” counterparts for the system of AcAla <sub>2</sub> NMe + Na <sup>+</sup> . . . . .	144
A.4	Original OPLS-AA FF parameters $\epsilon_{ij}$ and their “adjusted” counterparts for the system of AcAla <sub>2</sub> NMe + Na <sup>+</sup> using LASSO regression. . . . .	146
A.5	Original OPLS-AA FF parameters $V_1^{ij}$ , $V_2^{ij}$ , and $V_3^{ij}$ as well as their “adjusted” counterparts for the system of AcAla <sub>2</sub> NMe + Na <sup>+</sup> using LASSO regression. . . .	147



# List of Acronyms

<b>AM1</b>	Austin Model 1
<b>AMBER</b>	Assisted Model Building with Energy Refinement
<b>AMOEBA</b>	Atomic Multipole Optimized Energetics for Biomolecular Applications
<b>BSSE</b>	basis set superposition error
<b>CBS</b>	complete basis set
<b>CC</b>	coupled-cluster
<b>CCSD</b>	coupled-cluster method using Singles and Doubles excitation levels
<b>CCSD(T)</b>	coupled-cluster method using Singles, Doubles, and perturbative Triples excitation levels
<b>CCSDT</b>	coupled-cluster method using Singles, Doubles, and Triples excitation levels
<b>CHARMM</b>	Chemistry at Harvard Macromolecular Mechanics
<b>CI</b>	configuration interaction
<b>DFA</b>	density-functional approximation
<b>DFT</b>	density-functional theory
<b>DLPNO</b>	domain-based local pair natural orbital
<b>ESP</b>	electrostatic potential
<b>FF</b>	force field
<b>FFAFFURR</b>	Framework For Adjusting Force Fields Using Regularized Regression
<b>GGA</b>	generalized gradient approximation
<b>HF</b>	Hartree-Fock
<b>IR</b>	infrared
<b>LASSO</b>	least absolute shrinkage and selection operator
<b>LDA</b>	local-density approximation

## List of Acronyms

---

<b>MAE</b>	mean absolute error
<b>MBD</b>	many-body dispersion
<b>ME</b>	maximum error
<b>MNDO</b>	Modified Neglect of Diatomic Overlap
<b>MP2</b>	second-order Møller-Plesset perturbation theory
<b>NDDO</b>	Neglect of Diatomic Differential Overlap
<b>OPLS-AA</b>	Optimized Potentials for Liquid Simulations - All-Atom
<b>PES</b>	potential energy surface
<b>PM3</b>	Parametric Method 3
<b>PM6</b>	Parametric Method 6
<b>PM7</b>	Parametric Method 7
<b>TS</b>	Tkatchenko-Scheffler
<b>UV</b>	ultraviolet
<b>vdW</b>	van der Waals
<b>xc</b>	exchange-correlation
<b>ZDO</b>	Zero Differential Overlap
<b>ZORA</b>	zeroth order regular approximation
<b>ZPE</b>	zero-point energy

# 1 Motivation and Overview

This introductory chapter will give an overview on the three parts of research work this thesis contains, while at the same time highlighting the motivation that lead to tackling the specific topics. Hence, a more “traditional” introductory section will precede the respective research Chapters 3, 4, and 5, including detailed context and specific objectives.

Metal cations are essential for life, as approximately one third of the proteins in the human body require a metal cofactor for biological function [1, 2]. They often play an important role in shaping the three-dimensional structure of proteins and peptides. Furthermore, their presence may significantly influence important properties, *e.g.* binding sites, catalytic properties, and biological functions. As an example, it is hypothesized that protein misfolding of Alzheimer’s A $\beta$ -amyloid peptides into aggregated senile plaques inside the human brain of Alzheimer patients is promoted by metal ions such as zinc (Zn<sup>2+</sup>) [3]. Figure 1.1(a) shows the structure of the A $\beta$ (1–16)-Zn<sup>2+</sup> complex in aqueous solution at pH 6.5, determined from nuclear magnetic resonance (NMR) data [4]. One glutamic acid (Glu) residue and three histidine (His) residues act as ligands and tetrahedrally coordinate the zinc cation. Zinc ions are furthermore required for the catalytic function of more than 200 enzymes [5], an example being carbonic anhydrase [6]. Figure 1.1(b) shows the active site of human carbonic anhydrase II, determined by X-ray crystallography at 2.0 Å resolution [7]. Again, three His residues act as ligands to the central zinc ion of the active site.

These are but two examples where structures of protein-cation complexes have been determined experimentally. Besides necessary excellent knowledge of the experiment, it goes without saying that it is also very much desirable to have a very good fundamental and detailed theoretical understanding of the cation-peptide interaction systems. If both apply, the combination of experimental techniques with molecular simulations allows for structure elucidation as it helps to interpret experimentally obtained spectra. On the other side, a rigorous experiment-theory comparison allows for the assessment of the accuracy and predictive power of simulation approaches. Moreover, there may exist cases where a correct interpretation of both experimental and theoretical findings will not be possible using one without the other. After having introduced the experimental and theoretical background of the methods

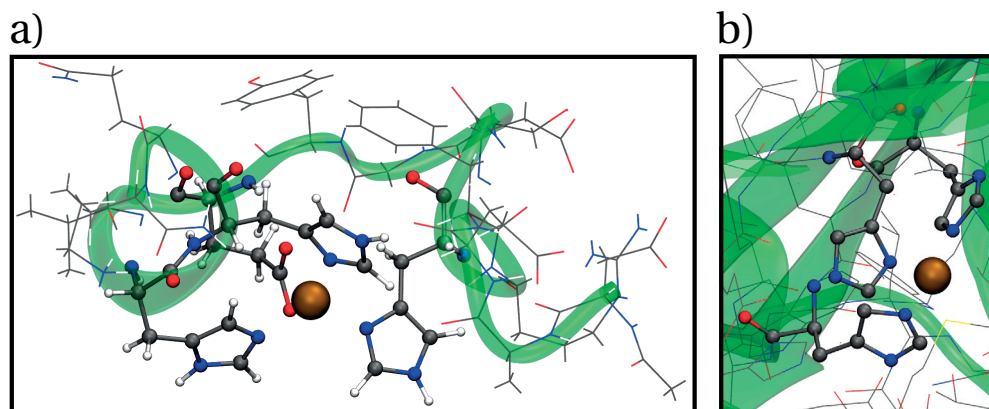


Figure 1.1 – Two examples of peptide-cation interaction sites: (a) Structure of the A $\beta$ (1–16)-Zn<sup>2+</sup> complex in aqueous solution at pH 6.5, determined from NMR data [4] (PDB ID: 1ZE9). (b) Active site of human carbonic anhydrase II, determined by X-ray crystallography at 2.0 Å resolution [7] (PDB ID: 1CA2). Images were created using VMD [8].

and techniques in Chapter 2 that are employed in this thesis, such an instant is presented in Chapter 3: There, the peptide AcPheAla<sub>5</sub>LysH<sup>+</sup> is investigated, a model system for studying helix formation in the gas phase, in order to fully understand the forces that stabilize the helical structure. In particular, the question of whether the local fixation of the positive charge at the peptide's C-terminus is a prerequisite for forming helices is addressed by replacing the protonated C-terminal lysine (Lys) residue by alanine (Ala) and a sodium cation (Na<sup>+</sup>). For sodiated AcPheAla<sub>6</sub>, globular rather than helical structures are found. Interestingly, the global minimum structure from simulation is not present in the experiment. Only a rigorous theory-experiment comparison will allow for the interpretation that this is due to high barriers involved in re-arranging the peptide-cation interaction that ultimately result in kinetically trapped structures being observed in the experiment.

The conclusions drawn in Chapter 3 rely on sufficiently accurate conformational energy hierarchies and infrared (IR) spectra calculated using density-functional theory (DFT) [9, 10] with computationally costly hybrid exchange-correlation (xc) functionals applied. On the other hand, the sampling of the global conformational space of the system relies on force field (FF) models that are low in computational costs. This in part inspired a study presented in Chapter 4 where the goodness of commonly applied levels of theory, *i.e.* force fields (FFs), semi-empirical quantum chemistry methods, density-functional approximations (DFAs) using a variety of xc functionals, composite methods, and wavefunction-based methods are being assessed and evaluated with respect to benchmark-grade coupled-cluster calculations. The methods are tested for their energetic description of peptide-cation systems, with a strong focus on benchmark systems in the gas phase consisting of either a bare acetylhistidine (AcH) or in presence of a Zn<sup>2+</sup> cation. While the choice of AcH + Zn<sup>2+</sup> complexes as benchmark systems has certainly been motivated by their biochemical relevance as shown by the examples of metalloproteomics given in the beginning of this introduction, they are furthermore



---

computationally feasible due to their small size, even for high-level methods, yet provide a challenging structure because of the tautomeric form of the neutral imidazole ring and the additional cation in the system.

Conventional FF calculations are associated with low computational costs and are widely used for molecular dynamics simulations or conformational searches [11]. However, in Chapter 4 it is concluded that they are not reliable enough for an accurate energetic description of these peptide-cation systems within “chemical accuracy” of 1 kcal/mol. In general, there exists a large discrepancy between the description of the potential energy surface from FFs and higher-level methods, *e.g.* DFT, second-order Møller-Plesset perturbation theory (MP2), *etc.* Two reasons are commonly attributed to this discrepancy: For one, FFs are optimized for condensed-phase systems instead of gas-phase systems with the latter being the main focus of this work due to the offered possibility of studying the “undamped” intramolecular interactions that shape peptides. Secondly, certain limitations in the FF description itself limit the accuracy of the energetic description. For example, as there are commonly no explicit bonds defined between cations and other atoms within the conventional empirical FF description, only non-bonded terms treating electrostatic and van der Waals interactions contribute to the overall empirical description of peptide-cation interactions. The work in Chapter 5 is described having two goals in mind: First, a machine-learning framework will be presented that serves as an interface between DFT and FF calculations. In essence, it serves to derive or “adjust” existing FF parameters from DFT calculations for a specific system in question, *e.g.* a particular peptide-cation system, using only a small number of structures for which single-point energy calculations are calculated at the DFT level. In contrast to conventional FF parameterization, this approach does not aim to yield general-purpose FF parameters but parameters adjusted for a specific system by the end-users themselves. Most importantly, torsional parameters or van der Waals parameters in the potential-energy function  $E_{\text{pot}}^{\text{FF}}$  of a particular FF (here: OPLS-AA [12–14]) are modified by fitting  $E_{\text{pot}}^{\text{FF}}$  against DFT energies using certain regularized linear regression models such as Ridge regression [15–17] or LASSO [18]. Secondly, because FF parameters are obtained from regularized regression methods using only energies calculated at the DFT level for a specific system in question, the set-up allows for immediate verification of how well the FF formulation itself is able to describe the potential energy, a venture to be undertaken quantitatively for the model systems AcAla<sub>2</sub>NMe and AcAla<sub>2</sub>NMe + Na<sup>+</sup>.



## **2 Theoretical and Experimental Background, Methods, and Techniques**

## 2.1 Amino Acids, Peptides, and Proteins

Peptides and proteins form one of four major classes of biomolecules, *i.e.* molecules present in organisms, with nucleic acids, lipids, and carbohydrates being the other three classes [19]. They are organic compounds that virtually affect every property that characterizes a living organism. To name but a few examples of their biological functions, the expression of genetic information encoded by nucleic acids depends almost entirely on proteins, they store and transport a variety of particles within organisms, they can act as hormones transmitting information between cells, or can act as enzymes increasing rates of chemical reactions that living organisms make use of [20]. Although being extremely diverse in structure and properties, peptides and proteins in organisms are all the same type of linear oligomer, being made of only 20 DNA encoded amino acids [21]. The various combinations of the same 20 amino acids, their chemical diversity, and the resulting diversity of the three-dimensional structures are the reasons for this large functional diversity of peptides and proteins. Amino acids are fairly simple organic compounds containing an amino group ( $\text{NH}_2$ ) and a carboxylic acid group ( $-\text{COOH}$ ) [22]. All 20 DNA encoded amino acids are  $\alpha$ -amino acids, meaning that both the amino group and the carboxylic acid group are attached to the same central carbon atom, the  $\alpha$ -carbon. Figure 2.1 shows the general form of the 20 DNA encoded amino acids. 19 of the 20 natural amino acids have the general form given in Figure 2.1(a) where the R-group denotes the side chain differentiating the different amino acids. The structural formula of the one exception, proline, is given in Figure 2.1(b) where the side chain is bonded to the nitrogen atom of the amino group. The 20 natural amino acids are summarized in Table 2.1 listing the respective names, abbreviations, and side chains.

The orientation of the four connecting groups, *i.e.* the amino group, the carboxylic acid group, the side chain, and the hydrogen atom, with respect to the  $\alpha$ -carbon ( $\text{C}_\alpha$ ) that acts as the chiral center defines two possibilities for optically active isomers, commonly named L- and D-isomers, as exemplarily shown in Figure 2.1(c). The mirror image of an isomer is called an enantiomer and usually behaves identically in most chemical environments. With the exception of glycine and proline, natural DNA encoded amino acids have the same stereochemistry at the  $\text{C}_\alpha$  as they are L-amino acids. The reason for that is not entirely understood and a matter of ongoing research [23]. L-isomers will be used throughout in this

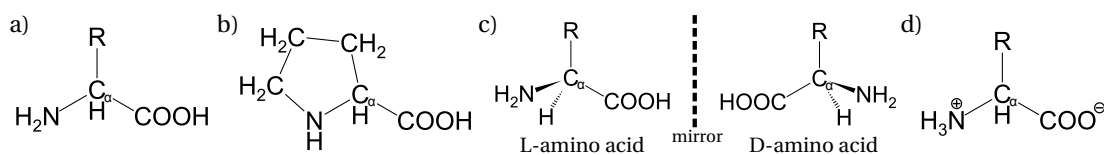
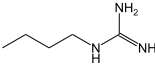



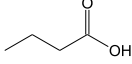
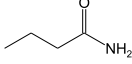



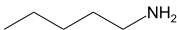
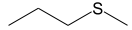

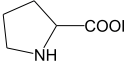


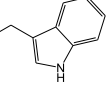
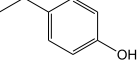



Figure 2.1 – (a) Structural formula of the general form of 19 of the 20 DNA encoded amino acids. The R-group denotes the side chain differentiating the different amino acids. The central carbon atom is commonly named the  $\alpha$ -carbon  $\text{C}_\alpha$ . (b) Structural formula of proline. (c) Depiction of the two theoretically possible optically active isomers of an amino acid with  $\text{C}_\alpha$  as the chiral center. (d) Zwitterion of the general form depicted in Figure (a).

## 2.1. Amino Acids, Peptides, and Proteins

Table 2.1 – Summary of the 20 natural amino acids listing the respective names, three- and one-letter abbreviations, and structural formulas of the side chain.

Name	Abbreviation		Side chain (R-group)
	3-letter	1-letter	
Alanine	Ala	A	$\text{—CH}_3$
Arginine	Arg	R	
Asparagine	Asn	N	
Aspartic acid	Asp	D	
Cysteine	Cys	C	
Glutamic acid	Glu	E	
Glutamine	Gln	Q	
Glycine	Gly	G	$\text{—H}$
Histidine	His	H	
Isoleucine	Ile	I	
Leucine	Leu	L	
Lysine	Lys	K	
Methionine	Met	M	
Phenylalanine	Phe	F	
Proline	Pro	P	 (drawn in full)
Serine	Ser	S	
Threonine	Thr	T	
Tryptophan	Trp	W	
Tyrosine	Tyr	Y	
Valine	Val	V	

## Chapter 2. Theoretical and Experimental Background, Methods, and Techniques

---

work.

Because of the present amino group and carboxylic acid group that can be (de)protonated, isolated amino acids or termini of peptides carry a basic and acidic component with them. Obviously, the degree of (de)protonation influences the physical properties of the amino acids and depends on the chemical environment. For example, in aqueous solution at neutral  $pH$ , *i.e.*  $pH \approx 7$ , the form that has both functional groups charged is the dominant species. A depiction for that is provided in Figure 2.1(d) where the termini have become charged to form a  $NH_3^+$  and a  $COO^-$  group.

The 20 natural amino acids possess a variety of chemical properties depending on their form, their combination of sequences in molecules, and the chemical environment. In the following, general properties of side chains are very briefly summarized for four amino acid residues that will appear numerous times throughout this work: Ala, Lys, Phe, and His. The Ala side chain consists of a methyl group ( $-CH_3$ ) and is therefore aliphatic, *i.e.* nonpolar and hydrophobic, meaning it is not interacting favorably with water but with other nonpolar atoms. The Lys side chain consists of a hydrophobic chain of four methylene groups capped by an amino group ( $-(CH_2)_4-NH_2$ ). The amino group of the side chain is able to participate in a multitude of reactions, is protonated and therefore positively charged under most physiological conditions. The Phe side chain consists of a benzyl group and therefore belongs to the aromatic side chains that allow for ultraviolet (UV) absorbance and fluorescence [24]. It is nonpolar and not chemically reactive under normal conditions applicable to proteins. The spectral properties of the residue are very sensitive to its immediate environment, thus allowing it to be used as a structural probe of protein structure [25]. Finally, the His side chain consists of an imidazole side chain that has a  $pK_a$  value of approximately  $6 < pK_a < 7$  [26], meaning both acid and base forms are present at neutral  $pH$ . The acid form with the imidazole ring protonated at both nitrogen atoms with its two equivalent contributing forms is shown in Figure 2.2(a). The positive charge is shared by both nitrogen atoms by resonance. The corresponding conjugate form of the neutral imidazole ring is shown in Figure 2.2(b). It exists as two tautomeric forms with the hydrogen atom on either the  $N_{\delta 1}$  or the  $N_{\delta 2}$  atom. The position of the hydrogen atom heavily depends on the local environment and both forms are present at neutral  $pH$ . The reactive amine can act as an effective nucleophilic catalyst. The nitrogen atom without the hydrogen is nucleophilic and an acceptor for hydrogen bonding, while the nitrogen atom with the hydrogen is electrophilic and a donor for hydrogen bonding, making this side chain very versatile [27].

Peptides and proteins are formally created when covalently linking amino acids together by peptide bonds. This process is called condensation and is depicted in Figure 2.3 for an example of two amino acids. The resulting dipeptide contains a terminus with an amino group and a terminus with a carboxylic acid group, commonly named N- and C-terminus, respectively. The biosynthesis of peptides and proteins always starts at the N-terminus, hence the amino sequence is always given from N- to the C-terminus. Protein biosynthesis takes place inside cells and denotes the last step in the process of gene expression where information

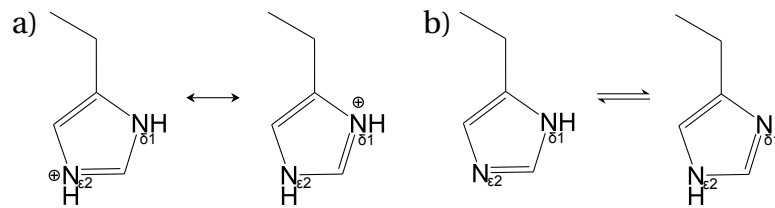


Figure 2.2 – Ionic forms of the histidine side chain at physiologically relevant pH. (a) Resonance hybrid forms of the ionized imidazole side chain. The two forms represent one structure as the positive charge is shared by the  $N_{\delta 1}$  and  $N_{\epsilon 2}$  atoms. (b) The two equivalent tautomeric forms of the non-ionized imidazole side chain.

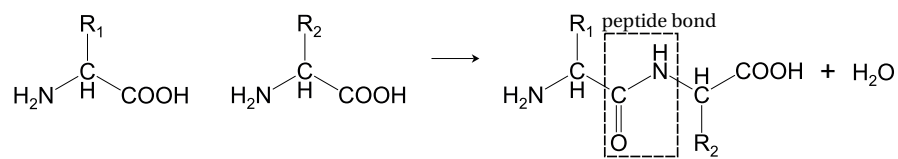


Figure 2.3 – Depiction of the condensation of two amino acids to form a peptide bond.

from a gene is used in the synthesis of proteins [28]. In short, DNA is transcribed into RNA inside the cell nucleus. The translation of this information into formed proteins – one amino acid after the other – then takes place in the cytoplasm of the cell and is undertaken by the ribosome. The large variety of proteins inherits from the large number of possibilities of the combination of amino acid sequences. Amino acids that are part of a peptide and proteins are referred to as residues. Different peptides or proteins differ only in the number and sequence of their amino acid residues. In other words, the sequence of amino acid residues identifies a peptide or protein unambiguously. Although there is no strict definition, one commonly refers to a short chain of amino acid residues with a defined sequence as peptide. Molecules that contain more than approximately 50 amino acid residues and possess a well-defined structure are denoted proteins. Medium-sized molecules with approximately 15 to 50 residues are sometimes referred to as polypeptides.

Figure 2.4 shows the schematic representation of an exemplary zwitterionic peptide consisting of five Ala residues, hence denoted Ala–Ala–Ala–Ala–Ala or shorter Ala<sub>5</sub>. The N-terminus, the C-terminus, the four peptide bonds, the five  $\alpha$ -carbons  $C_{\alpha}$ , and the five amino acid residues  $R_1, R_2, \dots, R_5$  are denoted. The linear chain consisting of the repeating sequence of the amide N, the  $C_{\alpha}$ , and the carbonyl C is called the backbone. Rotations around bonds are described as torsions or dihedral angles. A dihedral angle is defined as the angle between planes through two sets of three connecting atoms, having two atoms in common, and is taken to lie in the range  $-180^\circ$  to  $180^\circ$ . For example, in Figure 2.4, three backbone dihedral angles  $\phi$ ,  $\psi$ , and  $\omega$  are explicitly denoted. The dihedral angle around the peptide bond  $C(O)–N(H)$  is denoted  $\omega$ . The torsional angle around the bond  $C_{\alpha}–C(O)$  is denoted  $\psi$  and the torsional angle around the bond  $N(H)–C_{\alpha}$  is denoted  $\phi$ . The peptide bond dihedral angle can have two approximate values because of its partial double bond character resulting in a high rotational barrier [29]. If

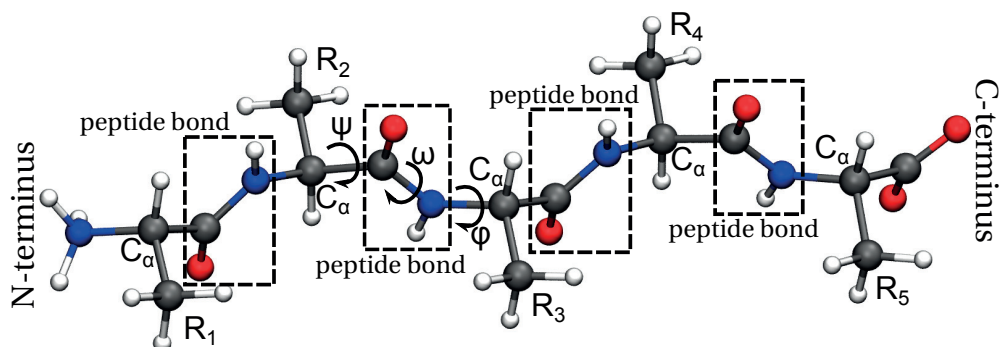


Figure 2.4 – Schematic representation of the exemplary zwitterionic peptide  $\text{Ala}_5$ . The N-terminus, the C-terminus, the four peptide bonds, the five  $\alpha$ -carbons  $C_\alpha$ , and the five amino acid residues  $R_1, R_2, \dots, R_5$  are denoted. Three specific examples of the backbone dihedral angles  $\phi$ ,  $\psi$ , and  $\omega$  are shown as well.

$\omega \approx 180^\circ$  ( $\approx -180^\circ$ ), meaning the chain is maximally extended (as in Figure 2.4), one commonly denotes such a configuration *trans*. For the other extreme case of  $\omega \approx 0^\circ$ , one commonly speaks of a *cis* configuration. Possible values of  $\phi$  and  $\psi$  are geometrically constrained by steric clashes of non-neighboring atoms and additional packing constraints [30].



## 2.2 The Structure and Energetics of Peptides in the Gas Phase

The ultimate goal of any research involving peptides or proteins is to understand their physical properties and biological functions. This requires to understand twofold: For one, the characterization of the underlying chemistry depends on the structure of the molecule itself, and secondly, the biological activity in addition depends on the interaction of the protein with its environment, *e.g.* water, membranes, other proteins, *etc.* [20], which in itself depends on environmental factors such as temperature, composition of the solvent, *pH* value, *etc.* In this work however, the focus lies (for the most part) on the former of the two aspects as the main goal is to study intramolecular interactions of peptides in the gas phase, *i.e.* in isolation. This is because gas-phase systems offer the opportunity to study the “undamped” intramolecular interactions that shape peptides, thereby shedding light on intrinsic structural motif propensities and bonding interactions.

The fundamental physical nature behind these interactions are rather well understood on an inter-atomic level. One thereby distinguishes between covalent and non-covalent interactions. A covalent bond [31], sometimes also called a molecular bond, is formed when involved atoms share electron pairs between them [32]. Covalent bonds are the strongest type of bonds in proteins and usually do not break during the lifetime of a protein [33]. Without explicitly highlighting it, proteins were discussed in Section 2.1 only in terms of their covalent structures. For example, covalent interactions are sufficient to describe the order of a sequence of amino acids inside a peptide because covalent bonds link the residues together. Obviously, the shapes of the side chains of the amino acid residues in a peptide create steric hindrance constraints influencing the forming and folding of the peptide. However, in order to accurately provide quantitative predictions of the overall three-dimensional structure of peptides and proteins, one also needs to accurately describe the physical nature of non-covalent interactions [34, 35], *i.e.* short-range repulsions, electrostatic forces, van der Waals interactions, and hydrogen bonds.

Short-range repulsions arise when two atoms approach each other and their electron orbitals begin to overlap. Following Pauli’s exclusion principle [36] that two identical electrons cannot occupy the same quantum state, this results in a strongly increasing repulsion. The corresponding repulsive energy arises steeply and is often described to scale with  $\sim r^{-12}$ , where  $r$  denotes the distance between the two atoms. The increase in energy is so steep that one often considers atoms as having definite occupying volumes that other atoms are unable to penetrate at normal temperatures. In fact, the schematic representation of the peptide shown in Figure 2.4 using a ball-and-stick representation already made use of this model. The radius of such a sphere of impenetrable volume around an atom is usually defined using the van der Waals radius [37], *i.e.* the distance of closest approach for another atom without forming a covalent bond. Different methods of determination exist [38], *e.g.* one of the more popular ones by Bondi [39] whose approach is based on a variety of experimental data like X-ray diffraction data and liquid state properties, among others. Van der Waals radii given in literature may vary, not only because of the missing strict definition, but also because they

## Chapter 2. Theoretical and Experimental Background, Methods, and Techniques

---

depend on the way an atom is covalently bonded [38].

Electrostatic forces between charges are the most fundamental non-covalent inter-atomic interactions. In vacuum, they are formally described by Coulomb's law where the energy of the electrostatic interaction  $E^{\text{Coulomb}}$  is given by

$$E^{\text{Coulomb}} = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j e^2}{r_{ij}}, \quad (2.1)$$

where  $\epsilon_0$  denotes the electric constant,  $e$  denotes the elementary charge, *i.e.* the magnitude of the electric charge carried by an electron,  $q_i$  and  $q_j$  denote the number of such charges on atoms  $i$  and  $j$ , respectively, and  $r_{ij}$  denotes the distance between the two atoms. Obviously, the simple form of Equation (2.1) is only valid when approximating the charges of atoms as point charges and neglecting finite sizes of ions. Furthermore, the localization of the electron density in peptides resulting in non-uniform distributions of negative and positive partial atomic charges results in electric dipoles, even if the peptide may have neutral net charge. Dipoles are formally described by its dipole moment  $\vec{\mu}$  given by

$$\vec{\mu} = q \vec{d}, \quad (2.2)$$

where  $q$  denotes the magnitude of the separated excess charge and  $\vec{d}$  denotes the distance vector between the two, directing from the negative charge towards the positive charge. The various electrostatic interactions between partial charges, permanent and induced dipoles on a number of atoms depend on each other and may result in fairly complex phenomena. Though in principle these kinds of interactions can always be described in terms of Coulomb's law given in Equation 2.1, it is often impractical to do so due to the complexity of a given system. Electrostatic interactions give rise to charge-charge interactions, *e.g.* ionic bonding that always includes some degree of covalent bonding [40], dipole-dipole interactions involving both permanent or induced dipoles, or more complex phenomena like cation- $\pi$  interaction, *i.e.* the interaction between the face of a  $\pi$  electron system of an aromatic ring such as the phenyl ring of the Phe side chain and an adjacent positively charged cation. The cation- $\pi$  interaction includes a substantial electrostatic component [41], as the  $\pi$  electrons in the aromatic ring are localized below and above the face of the ring, resulting in a partial negative charge in that region, opposed to a partial positive charge near the hydrogen atoms on its edge. The positive charge of a cation then creates a natural attraction towards the center of the face of the ring. However, other effects like polarization or charge-transfer may play a role as well for the complete understanding of the phenomenon [42].

Induced polarization effects between atoms and molecules are always present due to non-uniform distributions of partial atomic charges, as described in the last paragraph. This leads to weak attractive forces known as van der Waals (vdW) interactions that arise from three types of interactions [20]: Interactions between two permanent dipoles, those between a permanent and an induced dipole, and those between two mutually induced dipoles. The latter ones are known as London or dispersion interactions and are complex and quantum mechanical in

## 2.2. The Structure and Energetics of Peptides in the Gas Phase

---

nature [43,44]. In any case, all three components of van der Waals interactions scale with  $\sim r^{-6}$ , where  $r$  denotes the distance between two atoms, which is why they are often represented by an energy potential  $E^{\text{vdW}}$  including the attractive  $\sim r^{-6}$ -dependence as well as a short-range repulsion term discussed earlier with a steep  $\sim r^{-n}$ -dependence where  $n > 6$ :

$$E^{\text{vdW}} = \frac{C_n}{r^n} - \frac{C_6}{r^6} \quad (n > 6). \quad (2.3)$$

$C_n$  and  $C_6$  denote empirical constants. This form is called the Lennard-Jones potential [45]. In case of the common choice of  $n = 12$  the form is called the Lennard-Jones 12-6 (or 6-12) potential. The form obviously includes the approximation of the van der Waals interaction being independent of the orientation of interacting atoms or molecules. It furthermore assumes the vdW interactions to be occurring only pairwise between two atoms, neglecting many-body effects. The Lennard-Jones 12-6 potential is often expressed in its alternative form:

$$E^{\text{vdW}} = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right], \quad (2.4)$$

where  $\sigma$  and  $\epsilon$  are again empirical constants that relate to  $C_{12}$  and  $C_6$  by

$$C_{12} = 4\epsilon\sigma^{12}; \quad C_6 = 4\epsilon\sigma^6. \quad (2.5)$$

While weak in nature, the sum of all van der Waals interactions inside a molecule can add up to provide a significant stabilization to the three-dimensional structure of proteins [46].

Finally, hydrogen bonds [47] are formed when two electronegative atoms compete for the same hydrogen atom that is formally bonded to one of them, denoted the donor D, but also interacts favorably with the other, denoted the acceptor A. Although there are exceptions where the hydrogen atom is symmetrically centered between two electronegative atoms, it is usually covalently attached to one while also electrostatically interacting with the other ( $-D-H \cdots A-$ ). In peptides, hydrogen bonds frequently occur between the N-H and C=O groups of the peptide backbone, with the  $H \cdots O$  distance usually being  $\approx 1.9$  to  $2.0 \text{ \AA}$ . The predominant contribution to the hydrogen bond energy is of electrostatic nature, though an accurate quantum-chemical description requires to include exchange, polarization, and charge transfer contributions as well [48, 49]. Depending on the electronegativities of the donor and acceptor atoms, strengths and lengths of hydrogen bonds vary. Although hydrogen bonds are much weaker than covalent bonds, they have great significance in the structural properties of molecules [37]. This is also because of the important property of cooperativity of hydrogen bonds displayed in all classes of biological molecules [50]. Cooperativity, or non-additivity, thereby means the binding energy of a hydrogen bond structural system is greater than that of the sum of the individual bonds. In other words, the strength of the hydrogen bonds within a hydrogen bond chain is increased due to non-additive interactions between them arising from polarization and induced polarization effects [51].

The fundamental physical natures behind covalent and non-covalent interactions are formally sufficient to accurately describe the overall three-dimensional structure of peptides and

proteins in the gas phase. As laid out in the beginning of this section, the sequential order of amino acids of a peptide or protein is formally described by covalent interactions alone. On the other hand, in order to be able to describe specific structural features within the molecule, one needs to take into account non-covalent interactions as well and be able to describe them accurately and, if possible, on the same footing [35]. To reflect these different structural aspects, Linderstrøm-Lang classified the structures in the following way [52]: The primary structure means the sequential order of amino acids residues. The secondary structure refers to specific structural motifs or the geometric form of localized segments. The tertiary structure is the overall three-dimensional shape of the peptide or protein, usually composed of connected secondary structure elements. This structure classification is schematically represented in Figure 2.5.

In order to understand and identify secondary structure motifs in peptides and proteins, one needs to characterize peptide chain conformations. For one, this is done through means of the backbone dihedral angles  $\phi$ ,  $\psi$ , and  $\omega$ , as introduced in Section 2.1. As explained there, the dihedral angle  $\omega$  of the peptide bond only takes values around  $\omega \approx 0^\circ$  (cis configuration) or  $\omega \approx 180^\circ$  (trans configuration). On the other hand, possible values of  $\phi$  and  $\psi$  are only geometrically constrained by steric clashes of non-neighboring atoms and additional packing constraints [30]. In the gas phase, the actual three-dimensional conformation of a peptide chain is essentially determined by specific side-chain interactions and hydrogen bond patterns. It is thereby common that multiple hydrogen bonds are formed, resulting in a considerable stabilization of the secondary structure element due to the cooperativity effect explained above. There are three main secondary structure elements, namely helices,  $\beta$ -sheets, and turns, which are briefly described in the following. Helices and  $\beta$ -sheets are periodic secondary structure elements, meaning the torsional angles  $\phi$  and  $\psi$  of the associated consecutive amino acid residues have the same values. In contrast to that, turns are non-periodic secondary structure elements.

Helices are screw-like arrangements of the peptide backbone that are stabilized by intramolecular hydrogen bonds between the N–H and C=O groups of the peptide backbone. An example of a helix has already been shown in Figure 2.5(b) for the example of the Alzheimer’s disease amyloid  $\beta$ -peptide 1-16 region [4] where the helix is highlighted with a ribbon that is drawn through the backbone atoms. Keeping in mind that a typical H $\cdots$ O distance is usually  $\approx 1.9$  to  $2.0\text{Å}$ , hydrogen bonds in Figure 2.5(b) have been depicted with a dashed blue line when the H $\cdots$ O distance is smaller than  $3.0\text{Å}$ . There exist several helix types that can be characterized based on the intramolecular hydrogen bond patterns of the backbone alone, as depicted in Figure 2.6, namely  $\alpha$ -helices,  $3_{10}$ -helices, and  $\pi$ -helices. The  $\alpha$ -helix is the most common secondary structure element [54] and was originally proposed by Pauling *et al.* [55]. It comprises a right-handed spiral arrangement of the backbone with 3.6 amino acids residues per turn and the torsion angles  $(\phi, \psi) \approx (-57^\circ, -47^\circ)$  [29]. Stabilizing hydrogen bonds are directed backwards, *i.e.* from a C-terminal N–H group to a N-terminal C=O group, involving amino acid residues  $i + 4$  and  $i$ , hence the notation  $(\text{N–H})^{i+4} \rightarrow (\text{C=O})^i$ . Hydrogen bonds thereby form a “ring” consisting of 13 atoms. Taking into consideration the 3.6 amino acids per turn,

## 2.2. The Structure and Energetics of Peptides in the Gas Phase

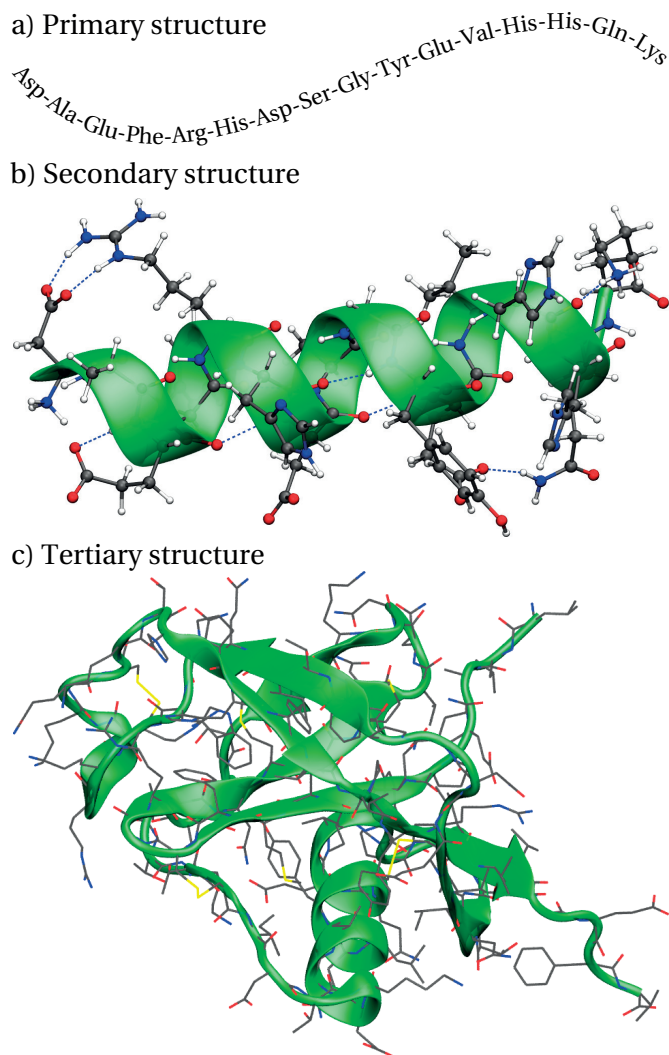


Figure 2.5 – Examples of primary, secondary, and tertiary structure of peptides and proteins. (a) Primary and (b) secondary structure of region 1-16 of the Alzheimer's disease amyloid  $\beta$ -peptide (PDB ID: 2BP4) [4]. (c) Tertiary structure of the N-terminal domain of the amyloid precursor protein (PDB ID: 1MWP) [53].

this gives rise to the formal alternative nomenclature of the  $\alpha$ -helix, the  $3.6_{13}$ -helix. The less common  $3_{10}$ -helix often caps an  $\alpha$ -helix in native peptides. As the nomenclature implies, it comprises a right-handed spiral arrangement of the backbone with 3 amino acids per turn and its involving hydrogen bonds consist of “rings” of 10 atoms. Similar to  $\alpha$ -helices, stabilizing hydrogen bonds in a  $3_{10}$ -helix are also directed backwards, but involve amino acid residues  $i + 3$  and  $i$ , hence the notation  $(\text{N-H})^{i+3} \rightarrow (\text{C=O})^i$  [56]. The recurring corresponding torsion angles are  $(\phi, \psi) \approx (60^\circ, -30^\circ)$ . The  $\pi$ -helix, or  $4.4_{16}$ -helix  $((\phi, \psi) \approx (-57^\circ, -70^\circ))$ , is rarely annotated despite occurring in 15% of known proteins [57]. Other helix types like polyproline helices exist and may occur commonly in proteins, especially when involving repeating proline residues [58]. More exotic helix types like  $2_7$ -type helices are formally possible but occur rarely

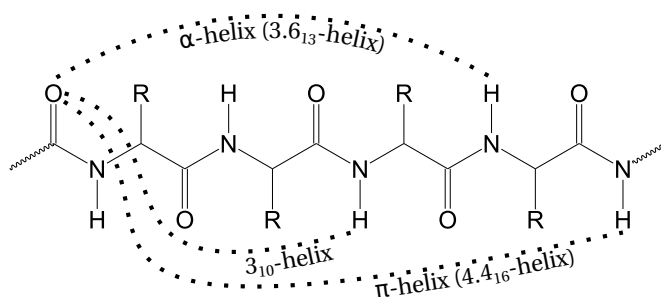


Figure 2.6 – Schematic view of hydrogen bond patterns in different helix types.

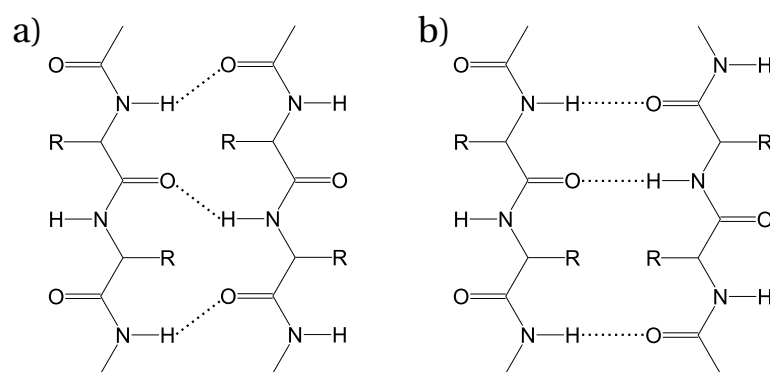


Figure 2.7 – Schematic view of hydrogen bond patterns in (a) parallel and (b) antiparallel  $\beta$ -sheets.

in native proteins [59].

The  $\beta$ -sheets, the second main common secondary structure elements, are hydrogen-bonded layer structures with hydrogen bonds being formed between two neighboring peptide chains and were originally proposed by Pauling *et al.* [60]. Two major  $\beta$ -sheet variants can be distinguished and are schematically presented in Figure 2.7: Parallel  $\beta$ -sheets are formally characterized by torsional angles  $(\phi, \psi) \approx (\pm 180^\circ, \pm 180^\circ)$  and two backbone chains being aligned in a parallel manner. Antiparallel  $\beta$ -sheets are characterized by two backbone chains being aligned in an antiparallel manner. Because present side chains distort the extended “zig-zag” conformation, antiparallel  $\beta$ -sheets formally display torsion angles  $(\phi, \psi) \approx (-139^\circ, 135^\circ)$  [29]. Several variants exist, *e.g.* twisted or backfolded forms, depending on the constitution of the sequence of amino acids inside a peptide.

While helices and sheets are unidirectional, loops reverse the direction of a peptide chain. Such loops are realized through turns that are often, but not necessarily, stabilized by a hydrogen bond. Depending on the number of amino acid residues involved, one classifies  $\alpha$ -,  $\beta$ -,  $\gamma$ -, and  $\pi$ -turns with five, four, three, and six amino acid residues involved, respectively [29]. Several different types of  $\alpha$ -turns can be classified [61] depending on the torsion angles of the three central amino acid residues. The most common class of turns are  $\beta$ -turns, or Venkatachalam-

## 2.2. The Structure and Energetics of Peptides in the Gas Phase

---

turns [62], where the distance of the  $C_{\alpha}$  atoms between amino acid residues  $i$  and  $i + 3$  is smaller than  $7 \text{ \AA}$  [63]. Again, several different types of  $\beta$ -turns can be classified depending on the characteristic torsional angles of the two central amino acid residues [64].

After having given an overview of the main secondary structure elements as well as the fundamental physical natures behind the interactions responsible for their formation, it is obvious that a formal description of peptide structure and dynamics is required in order to be able to explain and predict chemical and physical properties. In other words, one desires a concise description, both experimental and theoretical, of how a peptide folds into its native structure, a problem not yet solved and still a matter of ongoing research [65, 66]. More precisely, the problem to be solved actually requires twofold [67]: On one hand, one needs to predict the native three-dimensional structure of a given peptide system, and on the other hand, one wishes to describe the actual kinetics of the folding process. Obviously, solving the latter problem automatically includes solving the first one. In order to do so however, one needs to rely on directly folding a peptide chain, being it in experiment or theory, while the prediction of the native peptide structure can be based on the analysis of already known structures. The inherent problem when doing the former was already described by Levinthal in 1969 in what is called “Levinthal’s paradox” [68]: Assuming all peptide conformations were equally probable except for the native structure, meaning the native state can only be reached by an unbiased random search, this would lead to very large folding times. For example, a peptide with 30 amino acids, each of which can adopt 3 stable configurations, could be estimated to have  $3^{30}$  different configurations. Even if these configurations could be sampled at fastest possible time scales corresponding to vibrational modes of  $10^{-12} \text{ s}$ , it would still take  $\approx 6.5 \cdot 10^6$  years to do so [69]. The paradox then arises from the fact that proteins and peptides in living organisms arrive to their native form within timescales of less than a second [67]. Levinthal stated a solution to the paradox in that there were well-defined pathways to the native state [70], meaning the folding procedure was under “kinetic control” [71]. On the other hand, the “thermodynamic principle” by Anfinsen [72], also known as “thermodynamic hypothesis”, states that the native structure of a peptide is most favorable in thermodynamic terms, meaning the native structure corresponds to a kinetically accessible conformer with an overall reduction in free energy. This most importantly implies that native structure in a given environment is determined by the amino acid sequence of the peptide alone. The debate whether peptides reach their native structure following a specific pathway under kinetic control or in a pathway-independent manner under thermodynamic control, is ongoing [73]. Several theoretical models exist supporting one side or the other [73–75]. An approach in favor of the latter is described in the hypothesis of the existence of “folding funnels within free energy landscapes” [71, 76, 77] that postulates the folding of the peptide into the native state without the need for a definite pathway and is schematically depicted in Figure 2.8: In short, an unfolded peptide is high in both entropy and free energy. The free energy landscape means the free energy of each configuration as a function of the degrees of freedom of the system, *e.g.* torsional angles or other generic variables of the system. High entropy means a large number of possible configurations, and high free energy means the peptide is unstable

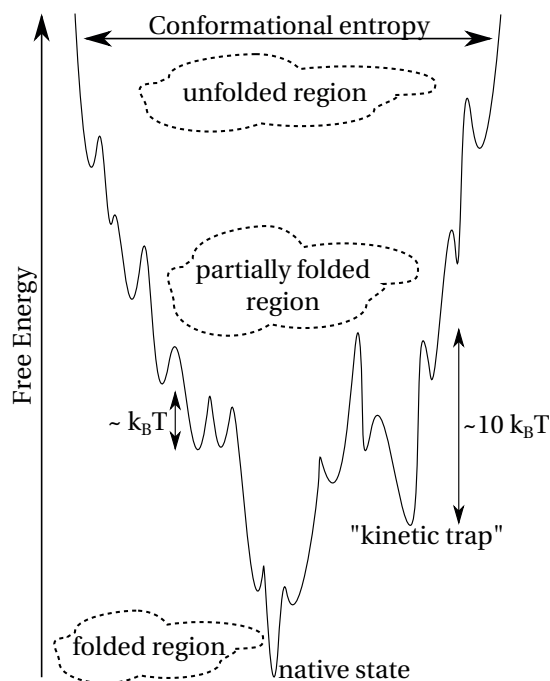


Figure 2.8 – Schematic depiction of a funnel-shaped free energy landscape. The “width of the funnel” indicates the amount of conformational entropy of the system. Local non-stable configurational states are surrounded by energy barriers of the order of  $\sim k_B T$ , while “kinetic traps” or intermediate states are surrounded by energy barriers significantly larger than that ( $\sim 10 k_B T$ ).

and thus being able to easily visit many different configurations. By following “down the funnel” of the free energy landscape the peptide folds, decreasing its free energy in the process. One appealing feature of this hypothesis is the inclusion of “kinetic traps” or intermediate states in its description: When “going down the funnel” partially folded peptides may become “trapped” in a local minimum of the free energy landscape higher in energy than that of the native structure and with deep surrounding energy barriers impossible for it to overcome, *i.e.* significantly larger than the order of  $\sim k_B T$  [78].

The work done within this thesis is primarily based on the assumption of correctness of the “folding funnel hypothesis”. As explained in the beginning of this section, the main goal of this thesis is to study intramolecular interactions of peptides in the gas phase. In order to do so and following the above assumption, this requires an accurate description and sampling of the free energy landscape, at least near the global minimum. Within the framework of this work, this will be aimed to achieve through – hopefully accurate – computer simulations of the potential energy surface (PES) [79] from which following quantities like free energies, vibrational modes, *etc.* are derived. The PES of a system is given by the potential energy as a function of all relevant atomic coordinates [69], and is thus a high-dimensional function even for small systems. A local minimum on the PES refers to a point from which a small displacement in either direction



## 2.2. The Structure and Energetics of Peptides in the Gas Phase

---

increases the potential energy. The lowest minimum is called the global minimum and usually refers to the native (folded) state of a peptide system. A detailed description on how to evaluate and sample the PES using vastly different theoretical models and levels of theory is provided in Section 2.3. In order to describe the thermodynamics of a real system, one needs to rely on the free energy surface that is a function of standard thermodynamic variables, *e.g.* temperature  $T$ , entropy  $S$ , pressure  $p$ , volume  $V$ , *etc.*, and is obtained from the PES by averaging over degrees of freedom of the system, *e.g.* torsional angles or other generic variables. This averaging provides an interpolation over the range for which the order parameters have physical meaning and thus provides a description that governs the behavior of a system in experiment. The work within this thesis relies on the Helmholtz free energy [80] for which free energy contributions are accounted for from internal degrees of freedom, consisting of vibrations and rotations, in addition to the potential energy on the PES. A detailed formulaic description is provided in Section 2.5. The Helmholtz free energy  $F$  is a natural function of its independent variables temperature  $T$  and volume  $V$ , and is formally defined by  $F = U - TS$ , with  $U$  and  $S$  denoting the internal energy and the entropy of the system, respectively [81]. It is related to the Gibbs free energy  $G$  [82] through  $G = U - TS + pV = F + pV$ , with pressure  $p$  and volume  $V$  denoting its natural independent variables. In biophysical experiments, the Helmholtz free energy is a useful quantity for experiments performed under conditions of constant temperature and volume, while the Gibbs free energy is a useful quantity for experiments performed under conditions of constant temperature and pressure [83]. As the goal of this work is to study peptide systems in the gas phase, *i.e.* in isolation, both experiment and theoretical calculations are essentially done at zero pressure, thus justifying the usage of the Helmholtz free energy for free energy contributions. Furthermore, throughout this work we are exclusively treating *relative* energies, *i.e.* comparing energy differences between different conformers (usually with respect to the global minimum) of the same system. Hence, the term containing the pressure, *i.e.* the  $pV$  term, cancels. In other words,  $\Delta G = \Delta F$ .

## 2.3 Description of the Potential Energy Surface (PES)

This section provides an overview on how to evaluate and sample the PES in computer simulations using vastly different theoretical models and levels of theory. This includes empirical force fields (FF), semi-empirical methods, density-functional theory (DFT), wavefunction-based methods like second-order Møller-Plesset perturbation theory (MP2), and coupled-cluster methods. All these methods have the common goal of describing the molecular system in place as accurate as possible while still being applicable from a computational point of view. Therefore, they may vastly differ in accuracy and certainly in computational costs. As a “rule of thumb”, *ab initio* (“first-principles”) methods, *i.e.* DFT methods and beyond, generally yield a higher predictive power in a wider range of problems, a consequence of them being in principle based entirely on the laws of quantum mechanics and not relying on experimental data other than the values of fundamental physical constants [84, 85]. This comes with the downside of them usually being much more computationally expensive, forcing the user to find a compromise between accuracy and computational costs for a specific task. After having given an overview of the different theoretical methods that will be made use of in this work, a subsection will be dedicated to the details of the applied computer simulations, see Subsection 2.3.9. Finally, Section 2.4 gives a brief overview on the sampling of the PES with a strong focus on the commonly applied method of basin-hopping in this work.

### 2.3.1 Force Fields

The empirical method of force fields (FFs) aims to provide an accurate description of structural properties of specific classes of systems [84]. It is based on the principle that these properties are primarily dictated by nearest-neighbor bonds. In essence, a bond between two atoms is to some extent assumed to be independent of which molecule it is a part of. Energetic variations are then ascribed to bond-angle contributions. Furthermore, higher-order contributions like van der Waals (vdW) and Coulomb interactions between non-bonded atoms are present as well and may be described similarly to their fundamental physical nature as laid out in the previous Section 2.2, in particular refer to Equations (2.1) and (2.4). Hence, the description of a FF is given by its potential energy  $E_{\text{pot}}^{\text{FF}}(\vec{R}^N)$  that is given as a function of positions  $\vec{R}_1, \dots, \vec{R}_N$  of the  $N$  nuclei of the system. In this classical approach, the potential energy  $E_{\text{pot}}^{\text{FF}}(\vec{R}^N)$  depends only on the nuclei positions and the types of atoms involved. It can be written as a sum of energy terms, each of them corresponding to qualitatively different interactions [84]:

$$E_{\text{pot}}^{\text{FF}}(\vec{R}^N) = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{tors}} + E_{\text{vdW}} + E_{\text{Coulomb}}, \quad (2.6)$$

where

$$E_{\text{bonded}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{tors}} \quad (2.7)$$

denotes the “bonded” contributions while

$$E_{\text{non-bonded}} = E_{\text{vdW}} + E_{\text{Coulomb}} \quad (2.8)$$

### 2.3. Description of the Potential Energy Surface (PES)

denotes the “non-bonded” contributions. For describing peptides, polypeptides, and proteins, examples of commonly applied conventional force fields are AMBER-99 [86, 87] (Assisted Model Building with Energy Refinement 99), CHARMM22 [88] (Chemistry at Harvard Macromolecular Mechanics 22), and OPLS-AA [12–14] (Optimized Potentials for Liquid Simulations - All-Atom) which are of similar form.

For the example of the OPLS-AA FF, the “bonded” terms are of the following form:

$$E_{\text{bonds}} = \sum_{i < j}^{1-2 \text{ atoms}} K_{ij}^r (r_{ij} - r_{ij}^0)^2, \quad (2.9)$$

$$E_{\text{angles}} = \sum_{i < j}^{1-3 \text{ atoms}} K_{ij}^\theta (\theta_{ij} - \theta_{ij}^0)^2, \quad (2.10)$$

$$E_{\text{tors}} = \sum_{i < j}^{1-4 \text{ atoms}} \left\{ \frac{V_1^{ij}}{2} (1 + \cos(\phi^{ij})) + \frac{V_2^{ij}}{2} (1 - \cos(2\phi^{ij})) + \frac{V_3^{ij}}{2} (1 + \cos(3\phi^{ij})) \right\}. \quad (2.11)$$

The sum in Equation (2.9) is over all pairs of atoms bonded to each other, also denoted as 1-2 atoms. The potential energy of the bonds is approximated as a harmonic oscillator, *i.e.* as a quadratic function of the displacement of the bond length  $r_{ij}$  from its reference length  $r_{ij}^0$ . The force constant  $K_{ij}^r$  and the reference length  $r_{ij}^0$  are empirical parameters taken from the AMBER FF that in turn were derived by fitting to structural and vibrational frequency data on small molecular fragments that make up proteins and nucleic acids [86]. In a similar fashion, the sum in Equation (2.10) is over all bond angles, *i.e.* atoms  $i$  and  $j$  that are separated by two bonds, also denoted as 1-3 atoms. The bond angle defined by the three atoms involved is denoted by  $\theta_{ij}$  and the reference bond angle is denoted by  $\theta_{ij}^0$ . The empirical parameters  $K_{ij}^\theta$  and  $\theta_{ij}^0$  are derived similarly as  $K_{ij}^r$  and  $r_{ij}^0$ . The sum in Equation (2.11) is over all torsional angles  $\phi^{ij}$ , *i.e.* atoms  $i$  and  $j$  that are separated by three bonds, also denoted as 1-4 atoms. The empirical parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  again depend on the atom classes of the four atoms defining the torsional angle. For the example of the OPLS-AA FF, they are derived from a least-squares fitting method using *ab initio* calculations [14]. For completeness, the “torsional” term of the potential energy  $E_{\text{tors}}$  in the description of the AMBER-99 and CHARMM22 FFs has a slightly different form:

$$E_{\text{tors}} = \sum_{i < j}^{1-4 \text{ atoms}} \sum_n \left\{ \frac{V_n^{ij}}{2} [1 + \cos(n\phi^{ij} - \phi_0^{ij})] \right\}. \quad (2.12)$$

Here,  $n$  denotes the number of minima over  $360^\circ$  of the torsional potential while the  $\phi_0^{ij}$  denote their location. The fitting methods for determining the empirical parameters  $V_n^{ij}$  and  $\phi_0^{ij}$  are described in References [87] and [88] for AMBER-99 and CHARMM22, respectively.

The “non-bonded” terms are intended to describe the fundamental non-covalent inter-atomic Coulomb and vdW interactions that have already been described in Section 2.2, compare to

Equations (2.1) and (2.4):

$$E_{\text{vdW}} = \sum_{i < j} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] f_{ij}, \quad (2.13)$$

$$E_{\text{Coulomb}} = \sum_{i < j} \frac{q_i q_j}{r_{ij}} f_{ij}. \quad (2.14)$$

The sum over all pairwise atomic Lennard-Jones and Coulomb contributions in Equations (2.13) and (2.14), respectively, runs over all pairs of atoms  $i$  and  $j$ . The corresponding 1-2 and 1-3 interactions are considered to be already implicitly included in their respective “bonded” contributions (Equations (2.9) and (2.10)). Within the description of the OPLS-AA FF, it was found to be necessary to scale the corresponding 1-4 interactions by a factor of  $\frac{1}{2}$  [13]. Hence, the scaling factor  $f_{ij}$  is given by

$$f_{ij} = \begin{cases} 0, & \text{for 1-2 and 1-3 atoms,} \\ \frac{1}{2}, & \text{for 1-4 atoms,} \\ 1, & \text{otherwise.} \end{cases} \quad (2.15)$$

The empirical parameters  $\epsilon_{ij}$ ,  $\sigma_{ij}$ , as well as the atomic partial charges  $q_i$  were derived from Monte Carlo simulations for pure liquids with the goal to reproduce the experimental heat of vaporization and molecular volume [89].

The functional form of the Coulomb term shown in Equation (2.14) includes one major limiting feature of conventional FFs, namely the inability to account for the influence of induced polarization and charge transfer which is due to the fixed atomic empirical partial charges  $q_i$ . In other words, the fixed form of  $E_{\text{Coulomb}}$  is incapable of describing the electric polarization, *i.e.* the redistribution of charge in space due to an electric field, being it for example an external macroscopic field or an induced electric field due to conformational changes of the peptide itself. Polarizable FFs aim to describe electronic polarization by including explicit models, *e.g.* the induced point dipole (IPD) model where point inducible dipoles  $\vec{\mu}_i$  are added to the  $N$  atomic sites of the molecule [90], the classical Drude oscillator model [91, 92], or the fluctuating charge (FQ) [92, 93] model. One example of this new generation of FFs is the AMOEBA [94–96] (Atomic Multipole Optimized Energetics for Biomolecular Applications) FF that is based on a similar potential energy form as conventional FFs but includes multipole representation of the fixed atomic partial charges and makes use of the IPD model. Its general functional form is given by

$$E_{\text{pot}}^{\text{AMOEBA}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{b}\theta} + E_{\text{oop}} + E_{\text{tors}} + E_{\text{vdW}} + E_{\text{elec}}^{\text{perm}} + E_{\text{elec}}^{\text{ind}}, \quad (2.16)$$

where the functional form of the “bonded” terms, *i.e.* bond stretching ( $E_{\text{bonds}}$ ), angle bending ( $E_{\text{angles}}$ ), and the coupling between the stretching and bending terms ( $E_{\text{b}\theta}$  and  $E_{\text{oop}}$ ) differs slightly when compared to the previously shown conventional force fields as they resemble the MM3 force field [97]. Similarly, the vdW term  $E_{\text{vdW}}$  adopts the buffered 14-7 functional

### 2.3. Description of the Potential Energy Surface (PES)

form [98] instead of the Lennard Jones 12-6 function in Equation (2.13). However, the major difference comes with the permanent electrostatic ( $E_{\text{elec}}^{\text{perm}}$ ) and induced electrostatic ( $E_{\text{elec}}^{\text{ind}}$ ) contributions. Concerning the latter term, in essence, one needs to describe the term by the scalar product of the induced dipole  $\vec{\mu}_i$  on the atomic site  $i$  with the permanent electric field  $\vec{E}^0(\vec{r}_i)$  due to the static charge distribution in the system, *i.e.*

$$E_{\text{elec}}^{\text{ind}} = -\frac{1}{2} \sum_{i=1}^N \vec{\mu}_i \cdot \vec{E}^0(\vec{r}_i), \quad (2.17)$$

where the induced dipole vector  $\vec{\mu}_i$  is expressed as

$$\vec{\mu}_i = \vec{\alpha}_i \left( \sum_{j \neq i} T_{ij}^1 \vec{M}_j + \sum_{k \neq i} T_{ik}^{11} \vec{M}_k \right). \quad (2.18)$$

$T_{ij}^1$  and  $T_{ik}^{11}$  hereby denote multipole-multipole and dipole-dipole interaction matrices, respectively. The permanent atomic multipole (PAM) vector  $\vec{M}_i$  at each atomic site  $i$  includes the corresponding charge, dipole, and quadrupole moments. The permanent electrostatic interaction energy  $E_{\text{elec}}^{\text{perm}}(r_{ij})$  between atomic sites  $i$  and  $j$  is expressed as

$$E_{\text{elec}}^{\text{perm}}(r_{ij}) = M_i^T T_{ij} M_j, \quad (2.19)$$

where  $T_{ij}$  denotes the interaction matrix between the two atomic sites. A detailed description of the terms is provided in Reference [96]. Within this work, two different parameterization versions of the AMOEBA FF will be used, namely AMOEBA-BIO09 [96, 99] and AMOEBA-PRO13 [100].

While the exact fitting procedure for obtaining empirical parameters varies between different FFs, they all have the common goal to reproduce certain features and properties by applying a fitting method using a certain selection of experimental or calculated *ab initio* data. Although the selection of the benchmark data for the commonly applied FFs tries to cover a broad range of features that allows for a rather general applicability, their performance will obviously be best for systems and configurations they were trained on. On the other hand, their reliability of quantitative predictions for systems different from those they were trained on is anything but clear and, in fact, can be misleading [101–104]. It should furthermore be emphasized, although obvious, that a FF treatment neglects any electronic effects. Nevertheless, FFs are widely used due to their cheap computational costs in comparison with *ab initio* methods. They are for example preferably applied for calculations of large systems or for sampling the large conformational space of peptides. However, for reliable quantitative predictions one commonly requires *ab initio* quantum-mechanical methods that will be discussed in the following subsections.

### 2.3.2 Schrödinger Equation and Born-Oppenheimer Approximation

From a quantum-mechanical point of view and within the scope of this work, a peptide system consisting of nuclei and electrons can formally be described by solving the non-relativistic time-independent Schrödinger equation [105]

$$\hat{H}\Psi = E\Psi, \quad (2.20)$$

where  $\hat{H}$  denotes the non-relativistic time-independent Hamilton operator,  $E$  denotes the total energy, and  $\Psi$  denotes the many-body wave function of the system. The Hamilton operator  $\hat{H}$  consists of five terms [106]:

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}_{n-n} + \hat{V}_{e-e} + \hat{V}_{n-e}. \quad (2.21)$$

The nuclear kinetic-energy operator  $\hat{T}_n$  is given by

$$\hat{T}_n = - \sum_{k=1}^M \frac{\hbar^2}{2M_k} \nabla_{\vec{R}_k}^2, \quad (2.22)$$

where the sum runs over all  $M$  nuclei that are assumed to be placed at the position  $\vec{R}_k$  and to have the mass  $M_k$ . The electronic kinetic-energy operator  $\hat{T}_e$  is given by

$$\hat{T}_e = - \sum_{i=1}^N \frac{\hbar^2}{2m_e} \nabla_{\vec{r}_i}^2, \quad (2.23)$$

where the sum runs over all  $N$  electrons that are assumed to be placed at the position  $\vec{r}_i$ . The electron mass is denoted by  $m_e$ . The potential energy operators of the system are simply described by the electrostatic energy due to charge interaction. Assuming the nuclear charges  $Z_k e$ ,  $k = 1 \dots M$ , the nucleus-nucleus potential-energy operator  $\hat{V}_{n-n}$  is then given by

$$\hat{V}_{n-n} = \frac{1}{2} \sum_{k_1 \neq k_2=1}^M \frac{1}{4\pi\epsilon_0} \frac{Z_{k_1} Z_{k_2} e^2}{|\vec{R}_{k_1} - \vec{R}_{k_2}|}. \quad (2.24)$$

All electrons have the same charge  $-e$ . Hence, the electron-electron potential-energy operator  $\hat{V}_{e-e}$  is given by

$$\hat{V}_{e-e} = \frac{1}{2} \sum_{i_1 \neq i_2=1}^N \frac{1}{4\pi\epsilon_0} \frac{e^2}{|\vec{r}_{i_1} - \vec{r}_{i_2}|}. \quad (2.25)$$

Finally, the nucleus-electron potential-energy operator  $\hat{V}_{n-e}$  is given by

$$\hat{V}_{n-e} = - \sum_{k=1}^M \sum_{i=1}^N \frac{1}{4\pi\epsilon_0} \frac{Z_k e^2}{|\vec{R}_k - \vec{r}_i|}. \quad (2.26)$$

For simplicity's sake, spin dependences have been neglected in the equations above: As fermions, two electrons are allowed to occupy any orbital, one with a  $\uparrow$ -spin and one with a  $\downarrow$ -spin. The position of each nucleus and electron is determined by three spatial coordinates

### 2.3. Description of the Potential Energy Surface (PES)

( $x$ ,  $y$ , and  $z$ ). Hence, even without explicitly considering the spin dependence, solving the Schrödinger equation means solving a problem of  $3M + 3N$  degrees of freedom for which the solution is not separable in its variables. Obviously, an exact solution is generally not possible and approximations must be made that are briefly discussed in the following.

The Schrödinger equation reads

$$[(\hat{T}_n + \hat{V}_{n-n}) + (\hat{T}_e + \hat{V}_{e-e} + \hat{V}_{n-e})] \Psi(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N) = E \Psi(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N). \quad (2.27)$$

The terms of the Hamilton operator have been re-grouped in a way that the first part depends solely on the nuclear coordinates  $\vec{R}_1, \dots, \vec{R}_M$ , whereas the latter part also depends on the electronic coordinates  $\vec{r}_1, \dots, \vec{r}_N$ . The Born-Oppenheimer approximation [107–109] relies on the fact that the mass of an electron is several thousand times smaller than that of a nucleus, *i.e.*

$$\frac{m_e}{M} \ll 1. \quad (2.28)$$

For example, even for the lightest nucleus, the proton, the ratio is [110]

$$\frac{m_e}{M_p} \approx \frac{1}{1836} \ll 1. \quad (2.29)$$

Hence, the electrons move much faster than the nuclei, meaning that for a given set of nuclear positions the electrons adjust their positions “immediately” with respect to the movement of the nuclei. The many-body wave function  $\Psi(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N)$  in Equation (2.27) can then be approximated as

$$\Psi(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N) = \Psi_n(\vec{R}_1, \dots, \vec{R}_M) \Psi_e(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N). \quad (2.30)$$

The separation of nuclear and electronic motions means in particular that the electronic wavefunction  $\Psi_e$  depends only parametrically on the nuclear coordinates  $\vec{R}_1, \dots, \vec{R}_M$ . Inserting Equation (2.30) into Equation (2.27), assuming terms of the form

$$-\frac{\hbar^2}{2M_k} \nabla_{\vec{R}_k}^2 \Psi_e(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N) \quad (2.31)$$

to be negligible, denoted the adiabatic approximation [111], and neglecting the kinetic energy of the nuclei completely, yields for the total energy  $E$  of the system [106]:

$$\begin{aligned} E &= \hat{V}_{n-n} + E_e(\vec{R}_1, \dots, \vec{R}_M) \\ &= \frac{1}{2} \sum_{k_1 \neq k_2=1}^M \frac{1}{4\pi\epsilon_0} \frac{Z_{k_1} Z_{k_2} e^2}{|\vec{R}_{k_1} - \vec{R}_{k_2}|} + E_e(\vec{R}_1, \dots, \vec{R}_M). \end{aligned} \quad (2.32)$$

The nuclei are treated as classical particles that give rise to an electrostatic potential in which the electrons move:

$$V(\vec{r}) = \sum_{k=1}^M \frac{1}{4\pi\epsilon_0} \frac{Z_k e^2}{|\vec{R}_k - \vec{r}|} \quad (2.33)$$

## Chapter 2. Theoretical and Experimental Background, Methods, and Techniques

but otherwise their effects are ignored. Without explicitly stating it, this approximation has already been used within the description of FFs in Subsection 2.3.1 where the electronic energy is described using the potential energy function given in Equation (2.6). Within the description of quantum mechanics however, the electronic energy  $E_e(\vec{R}_1, \dots, \vec{R}_M)$  in Equation (2.32) is then obtained by solving the electronic Schrödinger equation:

$$\underbrace{(\hat{T}_e + \hat{V}_{e-e} + \hat{V}_{n-e})}_{= \hat{H}_e} \Psi_e(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N) = E_e(\vec{R}_1, \dots, \vec{R}_M) \Psi_e(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N), \quad (2.34)$$

or explicitly written using Equations (2.23), (2.25), and (2.26):

$$\underbrace{\left( -\sum_{i=1}^N \frac{\hbar^2}{2m_e} \nabla_{\vec{r}_i}^2 + \frac{1}{2} \sum_{i_1 \neq i_2=1}^N \frac{1}{4\pi\epsilon_0} \frac{e^2}{|\vec{r}_{i_1} - \vec{r}_{i_2}|} - \sum_{k=1}^M \sum_{i=1}^N \frac{1}{4\pi\epsilon_0} \frac{Z_k e^2}{|\vec{R}_k - \vec{r}_i|} \right)}_{= \hat{H}_e} \Psi_e(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N) = E_e(\vec{R}_1, \dots, \vec{R}_M) \Psi_e(\vec{R}_1, \dots, \vec{R}_M, \vec{r}_1, \dots, \vec{r}_N). \quad (2.35)$$

In order to simplify Equation (2.35) it is common practice to use natural units [112, 113], *i.e.*

$$\begin{aligned} \hbar &= 1, \\ m_e &= 1, \\ |e| &= 1, \\ 4\pi\epsilon_0 &= 1. \end{aligned} \quad (2.36)$$

When also omitting the explicit parametric dependence of  $\Psi_e$  on the nuclear coordinates  $\vec{R}_1, \dots, \vec{R}_M$ , Equation (2.35) then becomes:

$$\underbrace{\left( -\frac{1}{2} \sum_{i=1}^N \nabla_{\vec{r}_i}^2 + \frac{1}{2} \sum_{i_1 \neq i_2=1}^N \frac{1}{|\vec{r}_{i_1} - \vec{r}_{i_2}|} - \sum_{k=1}^M \sum_{i=1}^N \frac{Z_k}{|\vec{R}_k - \vec{r}_i|} \right)}_{= \hat{H}_e} \Psi_e(\vec{r}_1, \dots, \vec{r}_N) = E_e \Psi_e(\vec{r}_1, \dots, \vec{r}_N). \quad (2.37)$$

In the following subsections, different approximate solutions to the electronic Schrödinger equation will be laid out.

### 2.3.3 Variational Principle for the Ground State and Hartree-Fock Method

Considering the system in any state  $\Psi$ , the expectation value of the energy  $E$  of the system is quantum-mechanically given by

$$E[\Psi] = \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle}, \quad (2.38)$$



### 2.3. Description of the Potential Energy Surface (PES)

---

where

$$\langle \Psi | \hat{H} | \Psi \rangle = \int \Psi^* \hat{H} \Psi d\vec{r}^N \quad (2.39)$$

and

$$\langle \Psi | \Psi \rangle = \int \Psi^* \Psi d\vec{r}^N. \quad (2.40)$$

The minimum-energy principle states that such an expectation value for any wave-function  $\Psi$  is always greater or equal than the energy  $E_0$  of the ground state of the system, *i.e.*

$$E[\Psi] \geq E_0. \quad (2.41)$$

The simple proof is *e.g.* provided in Reference [114]. An expectation value  $E[\Psi]$  calculated with any guessed wave-function  $\Psi$  will therefore always provide an upper bound to the ground state energy  $E_0$ . In order to obtain the ground state  $\Psi_0$  with the corresponding energy  $E[\Psi_0] = E_0$ , one would then vary the guessed wave-function  $\Psi$  until the functional  $E[\Psi]$  is minimized, instead of solving the Schrödinger equation directly, *i.e.*

$$E_0 = \min_{\Psi} E[\Psi]. \quad (2.42)$$

In addition, by using the method of Lagrangian undetermined multipliers [115], it can always be guaranteed that the final wave-function  $\Psi$  will be normalized, *i.e.*

$$\langle \Psi | \Psi \rangle = \int \Psi^* \Psi d\vec{r}^N = 1. \quad (2.43)$$

The Hartree method [112, 116–118], the oldest and simplest method to obtain an approximate solution to the electronic Schrödinger equation in (2.37), makes use of the variational principle. It approximates the wave-function  $\Psi$  as a product of individual non-interacting electron orbitals  $\psi_i$ ,  $i = 1, \dots, N$ , *i.e.*

$$\Psi_e^H(\vec{r}_1, \dots, \vec{r}_N) = \psi(\vec{r}_1) \cdot \psi(\vec{r}_2) \cdot \dots \cdot \psi(\vec{r}_N). \quad (2.44)$$

Applying the variational principle means considering variations in  $\langle \Psi_e^H | \hat{H}_e | \Psi_e^H \rangle$  under the constraint that all single non-interacting electron orbitals  $\psi_i$  are orthonormal, *i.e.*

$$\langle \psi_i | \psi_j \rangle = \delta_{ij}. \quad (2.45)$$

However, the Hartree *Ansatz* in Equation (2.44) does not take the indistinguishability of electrons into account, thereby violating the Pauli principle [36]. In other words, when interchanging any two electrons the wavefunction  $\Psi_e$  ought to be anti-symmetric. Within the Hartree-Fock approximation [117, 119, 120],  $\Psi_e$  is described by the so-called Slater determinant

that fulfills this condition:

$$\Psi_e^{\text{HF}}(\vec{r}_1, \dots, \vec{r}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\vec{r}_1) & \psi_2(\vec{r}_1) & \dots & \psi_N(\vec{r}_1) \\ \psi_1(\vec{r}_2) & \psi_2(\vec{r}_2) & \dots & \psi_N(\vec{r}_2) \\ \dots & \dots & \dots & \dots \\ \psi_1(\vec{r}_N) & \psi_2(\vec{r}_N) & \dots & \psi_N(\vec{r}_N) \end{vmatrix}. \quad (2.46)$$

Although already mentioned, it should be pointed out that the spin dependences have been neglected throughout, but their implementation in the above equations is straightforward. The ground state energy  $E_e^{\text{HF}}$  is then found to be given by [114, 121]

$$E_e^{\text{HF}} = \langle \Psi_0^{\text{H}} | \hat{H}_e | \Psi_0^{\text{H}} \rangle = \sum_{i=1}^N H_i + \frac{1}{2} \sum_{i,j=1}^N (J_{ij} - K_{ij}), \quad (2.47)$$

where

$$H_i = \int \psi_i^*(\vec{r}) \left[ -\frac{1}{2} \nabla^2 - \sum_{k=1}^M \frac{Z_k}{|\vec{R}_k - \vec{r}|} \right] \psi_i(\vec{r}) d\vec{r}. \quad (2.48)$$

$J_{ij}$  is denoted the Coulomb integral and is given by

$$J_{ij} = \int \int \psi_i(\vec{r}) \psi_i^*(\vec{r}) \frac{1}{|\vec{r} - \vec{r}'|} \psi_j(\vec{r}') \psi_j^*(\vec{r}') d\vec{r} d\vec{r}'. \quad (2.49)$$

$K_{ij}$  is denoted the exchange integral and is given by

$$K_{ij} = \int \int \psi_i^*(\vec{r}) \psi_j(\vec{r}) \frac{1}{|\vec{r} - \vec{r}'|} \psi_i(\vec{r}') \psi_j^*(\vec{r}') d\vec{r} d\vec{r}'. \quad (2.50)$$

Note that  $J_{ij} \geq K_{ij} \geq 0$  and  $J_{ii} = K_{ii}$ . One can similarly write Equation (2.47) as

$$E_e^{\text{HF}} = \langle \Psi_0^{\text{H}} | \hat{H}_e | \Psi_0^{\text{H}} \rangle = \sum_{i=1}^N H_i + E_{\text{Hartree}} + E_{\text{x}}, \quad (2.51)$$

where the Hartree energy  $E_{\text{Hartree}}$  is given by

$$E_{\text{Hartree}} = \sum_{i<j=1}^N J_{ij}, \quad (2.52)$$

and the exchange energy  $E_{\text{x}}$  is given by

$$E_{\text{x}} = - \sum_{i<j=1}^N K_{ij}. \quad (2.53)$$

Minimizing Equation (2.47) under the constraint that all single non-interacting electron orbitals  $\psi_i$  are orthonormal (see Equation (2.45)) yields the Hartree-Fock differential equations

$$\hat{F} \psi_i(\vec{r}) = \sum_{j=1}^N \epsilon_{ij} \psi_j(\vec{r}), \quad (2.54)$$

### 2.3. Description of the Potential Energy Surface (PES)

---

where the Fock operator  $\hat{F}$  is given by

$$\hat{F} = -\frac{1}{2}\nabla^2 - \sum_{k=1}^M \frac{Z_k}{|\vec{R}_k - \vec{r}|} + \hat{j} - \hat{k}. \quad (2.55)$$

The Coulomb operator  $\hat{j}$  and the exchange operator  $\hat{k}$  act on an arbitrary function  $f(\vec{r})$  in such a way that

$$\hat{j}f(\vec{r}) = \sum_{i=1}^N \int \psi_i^*(\vec{r}') \psi_i(\vec{r}) \frac{1}{|\vec{r} - \vec{r}'|} f(\vec{r}') d\vec{r}' \quad (2.56)$$

and

$$\hat{k}f(\vec{r}) = \sum_{i=1}^N \int \psi_i^*(\vec{r}') f(\vec{r}') \frac{1}{|\vec{r} - \vec{r}'|} \psi_i(\vec{r}) d\vec{r}'. \quad (2.57)$$

The values  $\epsilon_{ij}$  in Equation (2.54) are the Lagrange multipliers associated with the constraints of Equation (2.45). Focusing on solutions where

$$\epsilon_{ij} = \delta_{ij} \epsilon_j, \quad (2.58)$$

the Hartree-Fock equations become

$$\hat{F}\psi_i(\vec{r}) = \sum_{j=1}^N \epsilon_j \psi_j(\vec{r}), \quad (2.59)$$

where the  $\epsilon_i$  are given by [114]

$$\epsilon_i = \langle \psi_i | \hat{F} | \psi_i \rangle = H_i + \sum_{j=1}^N (J_{ij} - K_{ij}). \quad (2.60)$$

In principle, the  $\epsilon_i$  denote “orbital energies” of the single non-interacting electron orbitals associated with them, although a physical meaning is of course questionable as the electrons themselves are not independent single particles like the Hartree *Ansatz* implies. Under the assumption of unchanged orbitals on ionization, *i.e.* removing one electron from the orbital  $\psi_i$ , Koopmans’ theorem [122] states that

$$\epsilon_i = -I_i, \quad (2.61)$$

where  $I_i$  denotes the associated ionization energy.

Solving the Hartree-Fock equations in (2.59) is a traditional eigenvalue problem. However, a solution for them has to be found self-consistently since the Fock operator  $\hat{F}$  depends on the solution itself (through the operators  $\hat{j}$  and  $\hat{k}$  that in turn depend on the orbitals  $\psi_i$ ). Starting with an initial guess of the orbitals  $\psi_i$ , one generates the Fock operator (Equation (2.55) through Equations (2.56) and (2.57)) leading to new orbitals by solving the Hartree-Fock equations in (2.59). The new orbitals are then used to generate the new Fock operator leading again to new orbitals by solving the Hartree-Fock equations, *etc.* The procedure is repeated

until input and output agree within a certain threshold.

It is obvious that an exact analytical solution of the complicated integro-differential Hartree-Fock equations in (2.59) is commonly unfeasible. One would much rather rely on numerical solutions carried out with computer programs. The scaling behavior of the associated computational costs are formally of the order of  $\mathcal{O}(N^4)$  [123, 124] which stems from the electron repulsion integrals (Equations (2.56) and (2.57)) involved. In practice, a often used measure for  $N$  is the size of the basis set, *i.e.* the set of basic functions that the wavefunctions are expanded in. For accurate results, it must be ensured that a chosen finite basis set must be large enough in order to reproduce a “complete” basis set, commonly denoted as the complete basis set (CBS) limit [125]. Reducing computational costs by reducing the number of involved integrals gave rise to another family of methods, the so-called semi-empirical methods, that will be briefly discussed in the following subsection.

### 2.3.4 Semi-Empirical Quantum Chemistry Methods

Semi-empirical quantum chemistry methods are based on the Hartree-Fock method, but follow a simplification strategy by making approximations for computationally demanding terms [126, 127]. In order to account for caused errors, empirical parameters are incorporated into the formalism and fitted against experimental data or high-level calculations [128]. All semi-empirical methods reduce complexity by considering only valence electrons explicitly. Core electrons are treated by scaling down the nuclear charge or by introducing functions that treat the Coulomb effects of core electrons and nuclei simultaneously. The basis set of the valence electrons, *i.e.* the number of functions the valence orbitals are represented in, is purposefully reduced to a minimal set, meaning many semi-empirical methods use only s- and p-type orbitals and the basis functions are commonly Slater type orbitals [129]. The most important approximation in semi-empirical methods is the Zero Differential Overlap (ZDO) [126] approximation that assumes all products of basis functions located on different atoms to be neglected. Assuming a molecular system with atoms  $A, B, \dots$  and denoting the orthonormal s- and p-type orbitals associated with atoms  $K, L = A, B, \dots$  as  $\psi_{i,K}, \psi_{j,L}, \dots$ , the ZDO approximation then means

$$\int \psi_{i,K}^*(\vec{r}) \psi_{j,L}(\vec{r}) d\vec{r} = 0 \quad \text{if } K \neq L. \quad (2.62)$$

Consequently, electronic overlap integrals of the form (compare to Equation (2.45))

$$\int \psi_{i,A}^*(\vec{r}) \psi_{j,B}(\vec{r}) d\vec{r} = \langle \psi_{i,A} | \psi_{j,B} \rangle \quad (2.63)$$

are approximated as

$$\langle \psi_{i,A} | \psi_{j,B} \rangle = \delta_{ij} \delta_{AB}, \quad (2.64)$$

### 2.3. Description of the Potential Energy Surface (PES)

and electron repulsion integrals of the form (compare to Equations (2.49) and (2.50))

$$\int \int \psi_{i,A}^*(\vec{r}) \psi_{j,B}(\vec{r}) \frac{1}{|\vec{r} - \vec{r}'|} \psi_{k,C}^*(\vec{r}') \psi_{l,D}(\vec{r}') d\vec{r} d\vec{r}' = \langle \psi_{i,A} \psi_{k,C} | \psi_{j,B} \psi_{l,D} \rangle \quad (2.65)$$

are approximated as

$$\langle \psi_{i,A} \psi_{k,C} | \psi_{j,B} \psi_{l,D} \rangle = \delta_{A,B} \delta_{C,D} \langle \psi_{i,A} \psi_{k,C} | \psi_{j,A} \psi_{l,C} \rangle. \quad (2.66)$$

Defining the one-electron operator  $\hat{h}$  as

$$\begin{aligned} \hat{h} &= -\frac{1}{2} \nabla^2 - \sum_K^{M_{\text{nuclei}}} \frac{\tilde{Z}_K}{|\vec{R}_K - \vec{r}|} \\ &= -\frac{1}{2} \nabla^2 - \sum_K^{M_{\text{nuclei}}} v_K, \end{aligned} \quad (2.67)$$

where  $\tilde{Z}_K$  denotes the reduced nuclear charge of atom  $K$  due to the core electrons, one-electron integrals of the form (compare to Equation (2.48))

$$\int \psi_{i,A}^*(\vec{r}) \hat{h} \psi_{j,B}(\vec{r}) d\vec{r} = \langle \psi_{i,A} | \hat{h} | \psi_{j,B} \rangle \quad (2.68)$$

are approximated within the Neglect of Diatomic Differential Overlap (NDDO) [130] method as [126]

$$\begin{aligned} \langle \psi_{i,A} | \hat{h} | \psi_{j,A} \rangle &= \delta_{ij} \langle \psi_{i,A} | -\frac{1}{2} \nabla^2 - v_A | \psi_{i,A} \rangle - \sum_{K(\neq A)}^{M_{\text{nuclei}}} \langle \psi_{i,A} | v_K | \psi_{j,A} \rangle, \\ \langle \psi_{i,A} | \hat{h} | \psi_{j,B} \rangle &= \langle \psi_{i,A} | -\frac{1}{2} \nabla^2 - v_A - v_B | \psi_{j,B} \rangle, \\ \langle \psi_{i,A} | v_C | \psi_{j,B} \rangle &= 0. \end{aligned} \quad (2.69)$$

Other different semi-empirical approximation schemes exist that mainly differ in the treatment of the electron repulsion integrals (Equation (2.65)). Examples are the Intermediate Neglect of Differential Overlap (INDO) [131] approximation and the Complete Neglect of Differential Overlap (CNDO) [130, 132] approximation which reduce these integrals to just two parameters.

In order to account for such approximations, the remaining integrals can be (i) calculated directly using the functional form of the basis functions, (ii) described using empirical parameters that are based on experimental data, or (iii) described using empirical parameters that are fitted against experimental data. A combination of methods (i) and (ii) is applied for the NDDO, INDO, and CNDO approximations while a combination of methods (ii) and (iii) is applied for so-called Dewar-type, or “modified”, methods [84]. Examples include different versions of the INDO-based Modified Intermediate Neglect of Differential Overlap (MINDO) [133, 134] approximation, as well as NDDO-based parameterizations like the Modified Neglect of Diatomic Overlap (MNDO) [135], the Austin Model 1 (AM1) [136], and the Parametric Method 3 (PM3) [137]. The latter three methods are similar but differ in the core-core repulsion treat-

ment and the parameterization process itself. Within their description, Equations (2.69) are furthermore approximated as

$$\begin{aligned} \langle \psi_{i,A} | \hat{h} | \psi_{j,A} \rangle &= \delta_{ij} \langle \psi_{i,A} | -\frac{1}{2} \nabla^2 - v_A | \psi_{i,A} \rangle - \sum_{K(\neq A)}^{M_{\text{nuclei}}} \tilde{Z}_K \langle \psi_{i,A} \psi_{i,A} | \psi_{j,A} \psi_{j,A} \rangle, \\ \langle \psi_{i,A} | \hat{h} | \psi_{j,B} \rangle &= \frac{1}{2} \langle \psi_{i,A} | \psi_{j,B} \rangle (\beta_{\psi_i} + \beta_{\psi_j}), \end{aligned} \quad (2.70)$$

where  $\tilde{Z}_K$  again denotes the reduced nuclear charge of atom  $K$ , and  $\beta_{\psi_i}$  and  $\beta_{\psi_j}$  denote two atomic “resonance” parameters. Note that the overlap integral  $\langle \psi_{i,A} | \psi_{j,B} \rangle$  is calculated explicitly which is inconsistent with the ZDO approximation (Equation (2.64)), hence the “modified” labeling. Using only s- and p-type orbitals, only five *one-center* electron repulsion integrals (Equation (2.66)) exist within the NDDO approximation, namely  $\langle ss|ss \rangle$ ,  $\langle sp|sp \rangle$ ,  $\langle ss|pp \rangle$ ,  $\langle pp|pp \rangle$ , and  $\langle pp'|pp' \rangle$  (where  $p \neq p'$ ), each of which is described as an empirical parameter that needs to be obtained from atomic spectra. When modeling the remaining 22 *two-center* electron repulsion integrals (Equation (2.66)) as interactions between multipoles, they can then be expressed in terms of the five one-center electron repulsion integrals and the internuclear distances [138]. Within the MNDO approximation, the core-core repulsion between atoms  $A$  and  $B$  is described by

$$V_{\text{nn}}^{\text{MNDO}}(A, B) = \tilde{Z}_A \tilde{Z}_B \langle s_A s_A | s_B s_B \rangle (1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}), \quad (2.71)$$

where  $R_{AB}$  denotes the internuclear distance between the atoms in units of Å, and  $\alpha_A$  and  $\alpha_B$  are again empirical parameters that are fitted against experimental data. For interactions involving the pairs of atoms O–H or N–H, Equation (2.71) is replaced by the following form:

$$V_{\text{nn}}^{\text{MNDO}}(A, H) = \tilde{Z}_A \tilde{Z}_H \langle s_A s_A | s_H s_H \rangle (1 + R_{AH} e^{-\alpha_A R_{AH}} + e^{-\alpha_H R_{AH}}), \quad (2.72)$$

where atom  $A$  represents either the O or the N atom. Within the AM1 approximation, the core-core repulsion of Equation (2.71) has been modified by adding Gaussian functions, *i.e.*

$$V_{\text{nn}}^{\text{AM1}} = V_{\text{nn}}^{\text{MNDO}}(A, B) + \frac{\tilde{Z}_A \tilde{Z}_B}{R_{AB}} \sum_k \left( a_{kA} e^{-b_{kA}(R_{AB}-c_{kA})^2} + a_{kB} e^{-b_{kB}(R_{AB}-c_{kB})^2} \right), \quad (2.73)$$

where  $k = 2, 3, 4$  depending on the atoms involved. The empirical atomic parameters  $a_k$ ,  $b_k$ , and  $c_k$  are again obtained by fitting against experimental data. The PM3 approximation is almost similar to the AM1 approximation except that Equation (2.73) contains only two Gaussian terms per atom. In addition, a different parameterization scheme was used. In particular, the one-center electron repulsion integral parameters were also fitted against molecular data instead of being obtained from atomic spectral data. Other modifications exist that introduce various modifications or additional approximations and use a more complete parameter optimization process. For example, the main feature of the PM6 [139] approximation consists of the introduction of core-core diatomic interaction parameters into

## 2.3. Description of the Potential Energy Surface (PES)

---

its formulation such that the core-core repulsion of Equation (2.71) has been modified to

$$V_{\text{nn}}^{\text{PM6}}(A, B) = \tilde{Z}_A \tilde{Z}_B \langle s_A s_A | s_B s_B \rangle (1 + x_{AB} e^{-\alpha_{AB}(R_{AB} + 0.0003 R_{AB}^6)}), \quad (2.74)$$

where  $x_{AB}$  and  $\alpha_{AB}$  denote empirical diatomic interaction parameters. With respect to PM3, the empirical core-core parameters thereby increase from approximately 70 atomic parameters to approximately 5000 diatomic parameters. On the other hand, this additional flexibility allows for reducing the number of Gaussian core-core terms in Equation (2.73) down to one per atom. Finally, the PM6-based reparameterized PM7 [140] method employs treatment of dispersion and hydrogen bonds by adding specific energy correction terms.

### 2.3.5 Density-Functional Theory

Subsections 2.3.3 and 2.3.4 described methods and approximations aimed at solving the electronic Schrödinger equation (see Equation (2.37))  $\hat{H}_e \Psi_e = E_e \Psi_e$ . In any case, the solution  $\Psi_e$  is an  $N$ -electron wave-function that depends on  $3N$  spatial coordinates and  $N$  spin coordinates, making it a very complex object to describe. Assuming having obtained the solution  $\Psi_e$ , it is formally possible to derive any experimental observable from it, although in practice the complexity of  $\Psi_e$  generally increases the effort considerably or makes it often plain impossible to derive an accurate enough solution in the first place. Density-functional theory (DFT) is an electronic-structure calculation method that has the remarkable property of allowing to replace the complicated  $N$ -electron wave-function  $\Psi_e(\vec{r}_1, \dots, \vec{r}_N)$  and the associated Schrödinger equation (see Equation (2.37)) with the electron density  $\rho(\vec{r})$  and its associated calculation scheme [114]. In other words, the object  $\Psi_e(\vec{r}_1, \dots, \vec{r}_N)$  that depends on  $3N$  spatial coordinates can formally be replaced by the object  $\rho(\vec{r})$  that depends only on 3 spatial coordinates. Early methods that implied the idea of treating  $\rho(\vec{r})$  instead of  $\Psi_e(\vec{r}_1, \dots, \vec{r}_N)$  include the Thomas-Fermi model [141, 142] and the  $X\alpha$  model by Slater [143, 144]. They were constructed as approximations to solving the electronic Schrödinger equation and were thus not derived as an exact theory. In 1964 however, the theorems by Hohenberg and Kohn [145] provided the theoretical footing of DFT and showed that it is formally possible to calculate any ground-state property through the means of the electron density  $\rho(\vec{r})$  alone. This implies in particular that one does not need to know the  $N$ -electron wave-function  $\Psi_e(\vec{r}_1, \dots, \vec{r}_N)$  if instead the electron density  $\rho(\vec{r})$  can be obtained directly.

Assuming a total Hamilton operator of the form

$$\hat{H}_e = -\frac{1}{2} \sum_{i=1}^N \nabla_{\vec{r}_i}^2 + \sum_{i=1}^N V_{\text{ext}}(\vec{r}_i) + \frac{1}{2} \sum_{i_1 \neq i_2=1}^N \frac{1}{|\vec{r}_{i_1} - \vec{r}_{i_2}|}, \quad (2.75)$$

where  $V_{\text{ext}}(\vec{r}_i)$  denotes the (unknown) external potential that electron  $i$  moves in, *e.g.* the

electrostatic potential of the  $M$  nuclei of the system given by

$$V_{\text{ext}}(\vec{r}_i) = - \sum_{k=1}^M \frac{Z_k}{|\vec{R}_k - \vec{r}_i|}, \quad (2.76)$$

and assuming a given ground state electron density  $\rho(\vec{r})$  that fulfills

$$N = \int \rho(\vec{r}) \, d\vec{r}, \quad (2.77)$$

the first Hohenberg-Kohn theorem then states that  $V_{\text{ext}}(\vec{r}_i)$  is uniquely specified, *i.e.* it is not possible to have two different external potentials  $V_{\text{ext}}(\vec{r}_i)$  for a given ground state electron density  $\rho(\vec{r})$ . This implies that  $\rho(\vec{r})$  uniquely specifies all terms in the Hamilton operator  $\hat{H}_e$  which of course formally determines  $\Psi_e(\vec{r}_1, \dots, \vec{r}_N)$  for the ground state which in turn formally determines any ground-state property. Although the theorem does not yet provide any practical use, it states that there formally exists a one-to-one mapping between the ground state electron density  $\rho(\vec{r})$  and any ground-state property which can then formally be written as a functional of  $\rho(\vec{r})$ , an example being the electronic energy  $E_e$ :

$$E_e = E_e[\rho]. \quad (2.78)$$

Assuming  $E_0$  being the ground-state energy and  $\rho_0(\vec{r})$  being the associated ground-state electron density such that

$$E_e[\rho_0] = E_0, \quad (2.79)$$

the second Hohenberg-Kohn theorem then states that for any trial electron density  $\tilde{\rho}(\vec{r})$  that fulfills

$$N = \int \tilde{\rho}(\vec{r}) \, d\vec{r}, \quad (2.80)$$

a variational principle for the density functionals holds in such a way that

$$E_e[\tilde{\rho}] \geq E_0 = E_e[\rho_0]. \quad (2.81)$$

This variational principle for the density functionals is equivalent to the one for wave-functions in Equation (2.41). Assuming the actual functional form of  $E_e[\rho]$  was known, one could insert approximate electron densities  $\tilde{\rho}$  and minimize  $E_e[\tilde{\rho}]$  in order to improve any calculation for the ground state.

The proofs of the two Hohenberg-Kohn theorems are fairly simple and can be found *e.g.* in References [106, 114, 146]. The Levy-Lieb [147–150] formulation extends the original proof by Hohenberg and Kohn and eliminates the restriction to non-degenerate ground states. Of course, the Hohenberg-Kohn theorems do not yet provide any practical use since the functional for the electronic energy  $E_e[\rho]$  (or any other ground-state property) is commonly not explicitly known. In 1965 however, Kohn and Sham [151] provided a practical scheme for determining ground-state properties from the electronic density, which will be briefly laid out in the following.



### 2.3. Description of the Potential Energy Surface (PES)

Although the actual form of  $E_e[\rho]$  is not explicitly known, it can be expressed as

$$\begin{aligned} E_e[\rho(\vec{r})] &= T[\rho(\vec{r})] + \int V_{\text{ext}}(\vec{r})\rho(\vec{r}) d\vec{r} + \frac{1}{2} \int \int \frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r} d\vec{r}' + \tilde{E}_{\text{xc}}[\rho(\vec{r})] \\ &= T[\rho(\vec{r})] + \underbrace{\int V_{\text{ext}}(\vec{r})\rho(\vec{r}) d\vec{r}}_{= E_{\text{ext}}[\rho(\vec{r})]} + \underbrace{\int V_{\text{C}}(\vec{r})\rho(\vec{r}) d\vec{r}}_{= E_{\text{C}}[\rho(\vec{r})]} + \tilde{E}_{\text{xc}}[\rho(\vec{r})], \end{aligned} \quad (2.82)$$

using the Coulomb potential (or often called Hartree potential)

$$V_{\text{C}}(\vec{r}) = \int \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r}' \quad (2.83)$$

and the external potential  $V_{\text{ext}}(\vec{r})$  (see Equations (2.75) and (2.76)). All terms in Equation (2.82) are written as functionals of the electronic density  $\rho(\vec{r})$ . The first term denotes the kinetic energy and is the equivalent to the first term in Equation (2.75). The second term denotes the interaction energy due to the external potential  $V_{\text{ext}}(\vec{r})$  and is the equivalent to the second term in Equation (2.75). The third term denotes the Coulomb interaction energy and is the equivalent to the third term in Equation (2.75). The fourth term includes all (unknown) exchange and correlation effects. Applying the variational principle of Equation (2.81) means minimizing  $E_e[\rho]$  under the constraint that any trial electron density  $\rho(\vec{r})$  does not change the total number of electrons (see Equation (2.80)) which can always be guaranteed by making use of the method of Lagrangian undetermined multipliers, as laid out in Subsection 2.3.3. This yields [106]

$$\mu = \frac{\delta T}{\delta \rho} + V_{\text{ext}}(\vec{r}) + V_{\text{C}}(\vec{r}) + \frac{\delta \tilde{E}_{\text{xc}}}{\delta \rho}, \quad (2.84)$$

where  $\mu$  denotes the Lagrange multiplier associated with the constraint of Equation (2.80). Kohn and Sham introduced a fictitious system of non-interacting electrons with the same electron density  $\rho$  and the same electronic energy  $E_e$  as the real system. Hence, they are assumed moving in some effective potential  $V_{\text{eff}}(\vec{r})$ . For this model system, the expression of the electronic energy, equivalent to Equation (2.82), is much simpler and given by

$$E_e[\rho(\vec{r})] = \tilde{T}[\rho(\vec{r})] + \int V_{\text{eff}}(\vec{r})\rho(\vec{r}) d\vec{r}. \quad (2.85)$$

Note that  $\tilde{T}[\rho(\vec{r})] \neq T[\rho(\vec{r})]$ . Repeating the same procedure as before, this yields

$$\mu = \frac{\delta \tilde{T}}{\delta \rho} + V_{\text{eff}}(\vec{r}). \quad (2.86)$$

Comparing Equations (2.84) and (2.86) yields

$$\begin{aligned} V_{\text{eff}}(\vec{r}) &= \frac{\delta T}{\delta \rho} - \frac{\delta \tilde{T}}{\delta \rho} + V_{\text{ext}}(\vec{r}) + V_{\text{C}}(\vec{r}) + \frac{\delta \tilde{E}_{\text{xc}}}{\delta \rho} \\ &= V_{\text{ext}}(\vec{r}) + V_{\text{C}}(\vec{r}) + \frac{\delta E_{\text{xc}}}{\delta \rho}, \end{aligned} \quad (2.87)$$

where

$$E_{\text{xc}}[\rho(\vec{r})] = T[\rho(\vec{r})] - \tilde{T}[\rho(\vec{r})] + \tilde{E}_{\text{xc}}[\rho(\vec{r})] \quad (2.88)$$

denotes the exchange-correlation (xc) energy functional. Because the fictitious model system consists of non-interacting electrons, the Hamilton operator is of the simple form

$$\hat{H} = \sum_{i=1}^N \underbrace{\left[ -\frac{1}{2} \nabla_{\vec{r}_i}^2 + V_{\text{eff}}(\vec{r}_i) \right]}_{\hat{h}_{\text{eff}}}, \quad (2.89)$$

where  $\hat{h}_{\text{eff}}$  is a single-particle operator. The solution to the associated Schrödinger equation is yielded similarly as in Subsection 2.3.3 using a Slater determinant *Ansatz* (see Equation (2.46)) and is found self-consistently through the  $N$  single-particle equations, denoted the Kohn-Sham equations, that determine the single-particle orbitals  $\phi_i$ , denoted the Kohn-Sham orbitals:

$$\hat{h}_{\text{eff}}\phi_i = \epsilon_i\phi_i, \quad (2.90)$$

where

$$\rho(\vec{r}) = \sum_{i=1}^N |\phi_i(\vec{r})|^2. \quad (2.91)$$

While DFT in itself is an exact method, in practice approximations have to be made because the exact form of the xc functional  $E_{\text{xc}}[\rho(\vec{r})]$  (see Equation (2.88)) is commonly unknown. Although the actual form of  $E_{\text{xc}}[\rho(\vec{r})]$  should be very complex in general, it can often be approximated in a more or less reasonably simple manner. Obviously, a large variety of such density-functional approximations (DFAs) exist, commonly classified into different types depending on the features and formal properties of the xc functionals in question [152]. Such classifications are summarized in the following.

The **local-density approximation (LDA)** was already proposed by Kohn and Sham [151]. Within the LDA, the xc functional  $E_{\text{xc}}[\rho(\vec{r})]$  (see Equation (2.88)) is given by

$$E_{\text{xc}}^{\text{LDA}}[\rho] = \int \rho(\vec{r}) \epsilon_{\text{xc}}(\rho) d\vec{r}, \quad (2.92)$$

where the xc energy density  $\epsilon_{\text{xc}}(\rho)$  per particle is that of a uniform electron gas and thus not a local functional of  $\rho$ . Dividing  $\epsilon_{\text{xc}}(\rho)$  into exchange and correlation contributions such that

$$\epsilon_{\text{xc}}(\rho) = \epsilon_{\text{x}}(\rho) + \epsilon_{\text{c}}(\rho), \quad (2.93)$$

or equivalently

$$E_{\text{xc}}(\rho) = E_{\text{x}}(\rho) + E_{\text{c}}(\rho), \quad (2.94)$$

### 2.3. Description of the Potential Energy Surface (PES)

the exchange part can then be deduced analytically and is given by [114, 153]

$$\epsilon_x(\rho) = -\frac{3}{4} \left( \frac{3}{\pi} \right)^{1/3} \rho^{1/3}, \quad (2.95)$$

where  $\rho = \frac{3}{4\pi} \frac{1}{r_s^3}$ ,

and  $r_s$  denotes the radius of a sphere that contains one electron on average. The correlation energy density  $\epsilon_c(\rho)$  is not known analytically but accurate approximations exist, *e.g.* the PZ-LDA approximation by Perdew and Zunger [154], the PW-LDA approximation by Perdew and Wang [155], both based on quantum Monte Carlo results by Ceperley and Alder [156], and the VWN-LDA approximation by Vosko, Wilk, and Nusair [157]. The LDA is a good approximation for systems where the electron density is fairly uniform, *e.g.* bulk metals. It fails however for systems where the electron density has large variations, *e.g.* molecular systems with many hydrogen bonds, or weakly bound systems dominated by van der Waals interactions [158].

Within the description of the **generalized gradient approximation (GGA)** that generally reduces the typical error in LDA by a factor of 5 or more [159], gradients of the electron density are included in the xc functional as a variable, *i.e.* the xc functional is generally expressed in the form

$$E_{xc}^{GGA}[\rho] = \int \rho(\vec{r}) \epsilon_{xc}[\rho(\vec{r}), \nabla \rho(\vec{r})] d\vec{r}. \quad (2.96)$$

Widely used examples include the Perdew-Burke-Ernzerhof (PBE) [160] and the Becke-Lee-Yang-Parr (BLYP) [161, 162] xc functionals. For PBE, the xc functional is expressed as

$$E_{xc}^{PBE}[\rho] = E_x^{PBE}[\rho] + E_c^{PBE}[\rho], \quad (2.97)$$

where the exchange functional  $E_x^{PBE}[\rho]$  is given by

$$E_x^{PBE}[\rho] = \int \rho(\vec{r}) \epsilon_x^{LDA}[\rho(\vec{r})] F_x(s) d\vec{r}, \quad (2.98)$$

where  $\epsilon_x^{LDA}[\rho]$  is the exchange energy density in the uniform electron gas (see Equation (2.95)) and  $F_x(s)$  denotes the GGA enhancement factor depending on a dimensionless density gradient  $s$  which is defined as  $s = |\nabla \rho| / (2k_F \rho)$ , where  $k_F = (3\pi^2 \rho)^{1/3}$ . The enhancement factor  $F_x(s)$  is asked to satisfy a number of formal conditions and is expressed as

$$F_x(s) = 1 + \kappa - \frac{\kappa}{1 + \mu s^2 / \kappa}, \quad (2.99)$$

with  $\mu = \beta(\pi^2/3)$ ,  $\beta = 0.066725$ , and  $\kappa = 0.804$ . The PBE correlation functional  $E_c^{PBE}[\rho]$  is expressed as

$$E_c^{PBE}[\rho] = \int \rho(\vec{r}) [e_c^{LDA}(\rho, \xi) H(\rho, \xi, t)] d\vec{r}, \quad (2.100)$$

where  $e_c^{LDA}$  is the correlation energy density in PW-LDA approximation by Perdew and Wang [155],  $\xi = (\rho^\uparrow - \rho^\downarrow) / \rho$  denotes the relative spin polarization, and the function  $H(\rho, \xi, t)$  is

given by

$$H(\rho, \xi, t) = (e^2/a_0)\gamma\phi^3 \ln \left\{ 1 + \frac{\beta}{\gamma} t^2 \left[ \frac{1 + At^2}{1 + At^2 + A^2 t^4} \right] \right\}, \quad (2.101)$$

$$\text{with } A = \frac{\beta}{\gamma} \left[ e^{-\epsilon_c^{\text{LDA}}(\rho, \xi)/(\gamma\phi^3 e^2/a_0)} - 1 \right]^{-1},$$

where  $t = |\nabla\rho|/(2\phi k_s \rho)$  is a dimensionless density gradient,  $\phi(\xi) = [(1 + \xi)^{2/3} + (1 - \xi)^{2/3}]/2$  is a spin-scaling factor,  $k_s = (4k_F/\pi a_0)^{1/2}$ ,  $a_0 = \hbar^2/m_e e^2$ , and  $\gamma = (1 - \ln 2)/\pi^2$ . The PBE functional retains the correct features of LDA and includes inhomogeneity features that are supposed to be energetically important [158]. From a theoretical point of view, it does not contain empirical parameters obtained through fitting. The Becke '88 (B88) [161] exchange functional  $E_x^{\text{B88}}$  that makes up the exchange part of the BLYP xc functional  $E_{\text{xc}}^{\text{BLYP}}$  such that

$$E_{\text{xc}}^{\text{BLYP}}[\rho] = E_x^{\text{B88}}[\rho] + E_c^{\text{LYP}}[\rho], \quad (2.102)$$

contains only one empirical parameter  $\beta (= 0.0042)$  that is determined by a least-squares fit to exact atomic Hartree-Fock data obtained from six noble gas atoms. It is expressed as

$$E_x^{\text{B88}} = E_x^{\text{LDA}} - \underbrace{\beta \int \rho^{4/3} \frac{x^2}{1 + 6\beta x \sinh^{-1}(x)} d\vec{r}}_{= \Delta E_x^{\text{B88}}}, \quad (2.103)$$

where  $E_x^{\text{LDA}}$  denotes the LDA exchange functional and  $x = |\nabla\rho|/\rho^{4/3}$  is a dimensionless ratio. Unlike the other functionals, The Lee-Yang-Parr (LYP) [162] correlation functional  $E_c^{\text{LYP}}[\rho]$  is not based on the LDA but is instead derived from a correlation-energy formula due to Colle and Salvetti [163]. In its closed-shell form, it is given by

$$E_c^{\text{LYP}} = -a \int \frac{1}{1 + d\rho^{-1/3}} \left\{ \rho + b\rho^{-2/3} \left[ C_F \rho^{5/3} - 2t_w + \frac{1}{9} \left( t_w + \frac{1}{2} \nabla^2 \rho \right) \right] e^{-c\rho^{-1/3}} \right\} d\vec{r}, \quad (2.104)$$

$$\text{with } t_w = \frac{1}{8} \left( \frac{|\nabla\rho|^2}{\rho} - \nabla^2 \rho \right),$$

where  $C_F = \frac{3}{10}(3\pi^2)^{2/3}$ ,  $a = 0.04918$ ,  $b = 0.132$ ,  $c = 0.2533$ , and  $d = 0.349$ . In general, GGAs show improvements over LDAs in terms of binding energies, atomic energies, bond lengths, and angles [158].

A natural development after the GGAs consists of the inclusion of the Laplacian, *i.e.* the second derivative, of the electron density in the xc functional as a variable, beyond the electron density itself and its gradient. Such approximations of the xc functionals are denoted **meta-GGAs** and are thus generally expressed in the form

$$E_{\text{xc}}^{\text{meta-GGA}}[\rho] = \int \rho(\vec{r}) \epsilon_{\text{xc}}[\rho(\vec{r}), \nabla\rho(\vec{r}), \Delta\rho(\vec{r})] d\vec{r}. \quad (2.105)$$

Instead of being expressed in terms of  $\Delta\rho(\vec{r})$ , many meta-GGAs are expressed in terms of the

### 2.3. Description of the Potential Energy Surface (PES)

orbital kinetic energy density  $\tau(\vec{r})$  given by

$$\tau(\vec{r}) = \frac{1}{2} \sum_{i=1}^N |\nabla \phi_i(\vec{r})|^2, \quad (2.106)$$

where the  $\phi_i(\vec{r})$  denote the Kohn-Sham orbitals (see Equations (2.90) and (2.91)). Equation (2.105) then becomes

$$E_{xc}^{\text{meta-GGA}}[\rho] = \int \rho(\vec{r}) \epsilon_{xc}[\rho(\vec{r}), \nabla \rho(\vec{r}), \tau(\vec{r})] d\vec{r}. \quad (2.107)$$

The orbital kinetic energy density  $\tau(\vec{r})$  and  $\Delta\rho(\vec{r})$  are formally related:

$$\tau(\vec{r}) = -\frac{1}{2} \sum_{i=1}^N \phi_i^*(\vec{r}) \nabla^2 \phi_i(\vec{r}) + \frac{1}{4} \nabla^2 \rho(\vec{r}) \quad (2.108)$$

and Perdew and Constantin [164] presented evidence that both quantities carry essentially the same information beyond that is carried by  $\rho$  and  $\nabla\rho$ . Solving the Kohn-Sham equations (see Equations (2.90)) self-consistently, requires the evaluation of  $\delta E_{xc}/\delta\rho$ . However, since  $\tau(\vec{r})$  is not an explicit functional of the electron density  $\rho$ , this would in principle require some cumbersome methodological effort [165, 166]. In practice however, the derivative of  $E_{xc}^{\text{meta-GGA}}[\rho]$  is usually just evaluated with respect to the Kohn-Sham orbitals [167]. Examples of meta-GGA xc functionals include the group of Minnesota functionals developed by Truhlar and coworkers in Minnesota in 2005 and later, that are all parameterized against a broad range of chemical data [168]. Examples include the M06-L [169] xc functional designed for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and non-covalent interactions, and the M11-L [170] xc functional for transition metal thermochemistry, kinetics and non-covalent interactions. The latter xc functional incorporates a dual-range exchange strategy, meaning it makes use of two different meta-GGA functionals, one describing the short-range and one describing the long-range inter-electronic Coulomb interaction. Finally, the recently developed Strongly Constrained and Appropriately Normed (SCAN) [171] meta-GGA xc functional by Perdew and coworkers has been constructed to satisfy all 17 known possible exact constraints and further “appropriate norms” including the energies of rare-gas atoms and non-bonded interactions.

All previously mentioned local and semi-local approximations for the xc functional  $E_{xc}[\rho]$  suffer from the fact that non-locality is not fully taken into account, meaning that there must exist a formal accuracy limit that such calculations are able to reach. In particular, such approximations do not compensate entirely for the electronic self-interaction, *i.e.* the spurious interaction of an electron with itself. In the 1990s, based on the adiabatic connection approach [172], this motivated an advanced approach by Becke [173, 174]: While the correlation effects are still being treated within the DFT scheme, exchange effects are simultaneously treated using DFT and Hartree-Fock. In other words, the exchange part of the DFA xc functional is admixed with exact exchange from Hartree-Fock theory, resulting in so-called **hybrid exchange(-correlation) functionals**. The exact exchange from Hartree-Fock theory is thus

given by (compare to Equation (2.50))

$$E_x^{\text{exact}} = -\frac{1}{2} \sum_{i,j=1}^N \int \int \phi_i^*(\vec{r}) \phi_j(\vec{r}) \frac{1}{|\vec{r} - \vec{r}'|} \phi_i(\vec{r}') \phi_j^*(\vec{r}') d\vec{r} d\vec{r}', \quad (2.109)$$

where the  $\phi_i(\vec{r})$  again denote the Kohn-Sham orbitals. This approach introduces at least one empirical parameter, namely the so-called mixing parameters typically denoted  $\alpha_0, \alpha_1, \dots$  that regulate the relative proportions of exact exchange and DFAs. Hence, the optimum values of the mixing parameters depend on the physical parameters to which they are fitted [175]. For example, the widely popular B3LYP functional contains three parameters  $\alpha_0, \alpha_1, \alpha_2$  that were determined in order to accurately reproduce atomization energies, ionization potentials, proton affinities, and atomic energies of a given set of smaller molecules [174]. The corresponding xc functional is expressed as [176, 177]

$$E_{\text{xc}}^{\text{B3LYP}} = \alpha_0 E_x^{\text{exact}} + (1 - \alpha_0) E_x^{\text{LDA}} + \alpha_1 \Delta E_x^{\text{B88}} + (1 - \alpha_2) E_c^{\text{VWN}} + \alpha_2 E_c^{\text{LYP}}, \quad (2.110)$$

where  $\alpha_0 = 0.20$ ,  $\alpha_1 = 0.72$ ,  $\alpha_2 = 0.81$ ,  $E_x^{\text{exact}}$  denotes the exact exchange term given in Equation (2.109),  $E_x^{\text{LDA}}$  is the LDA exchange functional (see Equations (2.92) through (2.95)),  $\Delta E_x^{\text{B88}}$  denotes Becke's '88 gradient correction for exchange (second term in Equation (2.103)),  $E_c^{\text{VWN}}$  is the LDA correlation functional in VWN-LDA approximation by Vosko, Wilk, and Nusair [157], and  $E_c^{\text{LYP}}$  is the GGA correlation functional by Lee, Yang, and Parr given in Equation (2.104). Another example for a hybrid functional is given by the PBE0 [178, 179] model that contains one parameter  $\alpha_0 = 0.25$  which has been fixed *a priori* taking into account numerical results for molecular systems from fourth-order perturbation theory [180]. The corresponding xc functional is expressed as

$$E_{\text{xc}}^{\text{PBE0}} = \alpha_0 E_x^{\text{exact}} + (1 - \alpha_0) E_x^{\text{PBE}} + E_c^{\text{PBE}}, \quad (2.111)$$

where  $E_x^{\text{PBE}}$  and  $E_c^{\text{PBE}}$  denote the exchange and correlation parts of the GGA PBE xc functional given in Equations (2.97) through (2.101). The downside of this approach comes with increased computational costs as single-point energy evaluations using hybrid functionals are commonly at least one order of magnitude more expensive than their counterparts using (semi-)local functionals [181], although techniques to facilitate the treatment exist [182]. Other examples include hybrid meta-GGA xc functionals from the group of Minnesota functionals, *e.g.* M06 and M06-2X [183], two functionals incorporated with 27% and 54% exact exchange, respectively, M08-HX and M08-SO [184], two functionals incorporated with 52.23% and 56.79% exact exchange, respectively, and M11 [185], a hybrid meta-GGA functional with at least 42.8% exact exchange. While the M06 and M06-2X functionals are generally intended for overall good performance for chemistry, the latter functional is generally not suited for systems containing transition metals [168]. M08-HX is improved over M06-2X by making use of a cleaner functional form for both the exchange and the correlation parts which in turn became the base for the M11 functional. The M11 functional furthermore makes use of a long-range correction scheme [186], meaning the portion of exact exchange varies from 42.8%

### 2.3. Description of the Potential Energy Surface (PES)

at short-range to 100% at long-range. Finally, the meta-GGA hybrid xc functional SCAN0 [187] is constructed similarly as the PBE0 model shown in Equation (2.111), *i.e.*

$$E_{\text{xc}}^{\text{SCAN0}} = \alpha_0 E_{\text{x}}^{\text{exact}} + (1 - \alpha_0) E_{\text{x}}^{\text{SCAN}} + E_{\text{c}}^{\text{SCAN}}, \quad (2.112)$$

where again  $\alpha_0 = 0.25$ , and  $E_{\text{x}}^{\text{SCAN}}$  and  $E_{\text{c}}^{\text{SCAN}}$  denote the exchange and correlation parts of the meta-GGA SCAN xc functional.

Within the scheme of hybrid models where a non-local character gets introduced in the exchange part, properties governed by non-local correlation effects, *e.g.* non-covalent interactions, are commonly not described properly, which is due to the fact that the correlation part remains unchanged. The theoretical footing for improvement stems from the adiabatic connection formalism [172, 188–190] where  $E_{\text{xc}}[\rho]$  can be formally expressed as an integral of the form

$$E_{\text{xc}}[\rho] = \int_0^1 U_{\text{xc},\lambda}[\rho] d\lambda, \quad (2.113)$$

where the coupling-constant parameter  $\lambda$  regulates the assumed continuous adiabatic connection path between the fictitious non-interacting Kohn-Sham system ( $\lambda = 0$ ) and the physical system ( $\lambda = 1$ ) while all partially interacting systems ( $0 \leq \lambda \leq 1$ ) along the path maintain the same electron density  $\rho(\vec{r})$  as that of the physical system. Applying Görling-Levy second-order perturbation theory [191, 192] at the weakly interacting limit ( $\lambda \rightarrow 0$ ), the integrand  $U_{\text{xc},\lambda}[\rho]$  can then be formally expressed as

$$U_{\text{xc},\lambda}[\rho] \underset{\lambda \rightarrow 0}{\approx} E_{\text{x}}^{\text{exact}} + 2\lambda E_{\text{c}}^{\text{GL2}}, \quad (2.114)$$

where  $E_{\text{x}}^{\text{exact}}$  denotes the exact exchange given in Equation (2.109), and the second-order Görling-Levy correlation energy  $E_{\text{c}}^{\text{GL2}}$  may generally be well approximated [193] by only taking into account its double-excitation contributions  $E_{\text{c}}^{\text{PT2}}$  given by

$$E_{\text{c}}^{\text{PT2}} = \frac{1}{4} \sum_{i,j} \sum_{\alpha,\beta} \frac{\left| \int \int \phi_i^*(\vec{r}) \phi_j^*(\vec{r}') \frac{1}{|\vec{r}-\vec{r}'|} \phi_{\alpha}(\vec{r}) \phi_{\beta}(\vec{r}') d\vec{r} d\vec{r}' \right|^2}{\epsilon_i + \epsilon_j - \epsilon_{\alpha} - \epsilon_{\beta}}, \quad (2.115)$$

where  $\phi_i(\vec{r})$  and  $\phi_j(\vec{r})$  denote occupied Kohn-Sham orbitals, and  $\phi_{\alpha}(\vec{r})$  and  $\phi_{\beta}(\vec{r})$  denote unoccupied Kohn-Sham orbitals. The associated orbital energies are denoted  $\epsilon_i$ ,  $\epsilon_j$ ,  $\epsilon_{\alpha}$ , and  $\epsilon_{\beta}$ , respectively. Hence, expanding the idea of hybrid functionals by Becke [173] explained above, so-called **double hybrid functionals** were proposed [194–196] that not only included substitution of some portion of DFA exchange  $E_{\text{x}}^{\text{DFA}}$  by exact exchange  $E_{\text{x}}^{\text{exact}}$ , but also substitution of some portion of DFA correlation  $E_{\text{c}}^{\text{DFA}}$  by second-order perturbative correlation  $E_{\text{c}}^{\text{PT2}}$ , *i.e.*

$$E_{\text{xc}}^{\text{double-hybrid}} = (1 - \alpha_{\text{x}}) E_{\text{x}}^{\text{DFA}} + \alpha_{\text{x}} E_{\text{x}}^{\text{exact}} + (1 - \alpha_{\text{c}}) E_{\text{c}}^{\text{DFA}} + \alpha_{\text{c}} E_{\text{c}}^{\text{PT2}}, \quad (2.116)$$

where the scaling parameters  $\alpha_{\text{x}}$  and  $\alpha_{\text{c}}$  regulate the portions of substitution for exchange and correlation, respectively. Examples of double hybrid xc functionals include the XYG3 [197]

functional that aims for accurate descriptions of non-bonded interactions, thermochemistry, and thermochemical kinetics. It is of the form similar to the B3LYP xc functional shown in Equation (2.110), and expressed as

$$E_{xc}^{XYG3} = \alpha_0 E_x^{\text{exact}} + (1 - \alpha_0) E_x^{\text{LDA}} + \alpha_1 \Delta E_x^{\text{B88}} + (1 - \alpha_2) E_c^{\text{LYP}} + \alpha_2 E_c^{\text{PT2}}, \quad (2.117)$$

where the three mixing parameters  $\alpha_0 = 0.8033$ ,  $\alpha_1 = 0.2107$ , and  $\alpha_2 = 0.3211$  were determined empirically by fitting to thermochemical data [198]. All energy terms are thereby first evaluated using the Kohn-Sham orbitals, orbital energies, and associated electron density obtained from the B3LYP xc functional given in Equation (2.110).

Computational evaluation of the second-order perturbative correlation term in Equation (2.115) formally scales with  $\mathcal{O}(N^5)$ , where  $N$  denotes the size of the system, while regular hybrid DFA evaluations scale with  $\mathcal{O}(N^4)$  [123]. Obviously, this must imply certain computational limits for larger chemical systems. Recent so-called “low-cost” **composite electronic structure approaches** aim to (partly) overcome such limitations with a computationally more efficient methodology and without having to sacrifice accuracy. As an example, the PBEh-3c [199] scheme is based on the GGA PBE xc functional that has been modified into a hybrid functional with a relatively large amount of 42% of non-local exact exchange. The orbitals are expanded in computationally light Ahlrichs-type split valence double-zeta atomic orbital Gaussian basis sets [200]. In addition, Grimme’s empirical pairwise additive D3 correction method [201] (see Subsection 2.3.6) is applied in order to account for long-range dispersion. Finally, the third methodological correction consists of a global counterpoise-correction scheme [202] that accounts for the so-called basis set superposition error (BSSE) [203, 204], see Subsection 2.3.9 for details.

### 2.3.6 *A posteriori* van der Waals Correction Schemes in Density-Functional Theory and Semi-Empirical Quantum Chemistry Methods

With the exception of the latter two approaches presented in the last subsection, the “conventional” xc functionals in (semi-)local and hybrid approximation are not able to properly describe long-range electron correlation effects by design. In particular, this includes their inability to appropriately model long-range dispersion effects or van der Waals interactions [205]. The fundamental physical nature of dispersion effects have already been motivated in Section 2.2, refer in particular to Equation (2.3). Indeed, many systems containing biomolecules rely on van der Waals interaction treatments for an accurate energetic description [206, 207]. Following Equation (2.3), van der Waals interactions dominate in the long-range regime, *i.e.* in the region of negligible overlap between electronic charge densities of atomic fragments, and scale with  $\sim r^{-6}$ , where  $r$  denotes the distance between two atoms. In other words, its asymptotic behavior is described as

$$E^{\text{vdW}} \sim -\frac{C_6}{r^6}, \quad (2.118)$$



### 2.3. Description of the Potential Energy Surface (PES)

where  $C_6$  denotes the associated dispersion coefficient. One popular approach for dispersion-corrected DFAs consists of additive corrections, either pairwise in nature or also including many-body terms, which is for two reasons: Actual computational costs in evaluating these corrections are very little in comparison to the self-consistent Kohn-Sham calculation, and secondly, additive corrections may be combined with almost any (semi-)local DFA, as they are generally evaluated *a posteriori*, *i.e.* after the self-consistent Kohn-Sham treatment in DFT, and then simply added to the Kohn-Sham result [208], *i.e.*

$$E^{\text{DFA+vdW}} = E^{\text{DFA}} + E^{\text{vdW}}, \quad (2.119)$$

where  $E^{\text{DFA}}$  denotes the total energy of the system obtained in DFA and  $E^{\text{vdW}}$  denotes the dispersion correction commonly specific to the DFA. Following this approach, Equation (2.118) can then be generalized considering the multipolar expansion of the interatomic interactions [209]

$$E^{\text{vdW}} = E^{\text{vdW,(2)}} + E^{\text{vdW,(3)}} + \dots, \quad (2.120)$$

where  $E^{\text{vdW,(2)}}$  contains pairwise contributions,  $E^{\text{vdW,(3)}}$  contains three-body dispersion contributions, *etc.* The leading term considers all interactions between any pairs of atoms  $A$  and  $B$ , and is itself expressed in terms of not only the  $\sim r^{-6}$  contribution but also contains terms of order higher than the dipole-dipole interaction [210–212], *i.e.*

$$E^{\text{vdW,(2)}} = - \sum_{n=6,8,10,\dots} \frac{1}{2} \sum_{A \neq B} \frac{C_n^{AB}}{r_{AB}^n} f_n(r_{AB}). \quad (2.121)$$

The damping functions  $f_n$  thereby denote one-dimensional functions of the interatomic distance  $r_{AB}$  that fulfill

$$\begin{aligned} f_n(r_{AB}) &\xrightarrow{r_{AB} \rightarrow 0} 0 \\ \text{and } f_n(r_{AB}) &\xrightarrow{r_{AB} \rightarrow \infty} 1. \end{aligned} \quad (2.122)$$

For one, they serve to shut down the dispersion contribution at short range in order to avoid the singularity at  $r_{AB} \rightarrow 0$ , and secondly, they need to govern a seamless connection between the asymptotic long-range region and the short-range region that is mostly described by the underlying DFA. Furthermore, the flexible parameters of the damping functions may also be able to describe certain intermolecular interaction energies the underlying DFA is not able to reproduce [209]. Based on the above expressions, a variety of schemes and models exist for deriving the  $C_n$  coefficients with varying degrees of accuracy and empiricism. Examples of such dispersion correction schemes include the heavily parameterized and widely popular DFT-D3 model by Grimme and coworkers [201], as well as the parameter-free pairwise Tkatchenko-Scheffler van der Waals scheme ( $\text{vdW}^{\text{TS}}$ ) [213].

Within the description of the DFT-D3 model, the specific expression of Equation (2.121) is given by

$$E^{\text{D3,(2)}} = - \frac{1}{2} \sum_{A \neq B} \left[ \frac{C_6^{AB}}{r_{AB}^6} f_6(r_{AB}) + s_8 \frac{C_8^{AB}}{r_{AB}^8} f_8(r_{AB}) \right], \quad (2.123)$$

## Chapter 2. Theoretical and Experimental Background, Methods, and Techniques

where  $s_8$  is an adjustable parameter specific to each DFA xc functional and obtained by a fitting procedure using standard benchmark sets [201]. Higher-order contributions in Equation (2.123) have been found to make the method more unstable and are thus omitted. Different variations of the damping functions include the version by Becke and Johnson [214, 215], a “modified” version labeled D3M [216], and the so-called zero-damping function [217] that is primarily used in this work and given by

$$f_{d,n}(r_{AB}) = \frac{1}{1 + 6(r_{AB}/(s_{r,n}r_0^{AB}))^{-\alpha_n}}, \quad (\text{with } n = 6, 8), \quad (2.124)$$

where  $s_{r,n}$  is the order-dependent scaling factor of the cutoff radii  $r_0^{AB}$ . In particular,  $s_{r,n=8} = 1$ , and  $s_{r,n=6}$  is optimized using a least-squares fitting procedure. The cutoff radii  $r_0^{AB}$  were calculated from DFT calculations for all possible pairs of atoms  $A - B$ , resulting in 4465 values for 94 elements. Finally, the empirical parameters  $\alpha_n$  have been set to  $\alpha_6 = 14$  and  $\alpha_8 = 16$ . Based on the Casimir-Polder formalism [218], the dispersion coefficients  $C_n^{AB}$  in Equation (2.123) have been calculated using *ab initio* time-dependent (TD)DFT and employing recurrence formulas for the multipole terms in higher-order. The three-body dispersion contributions in Equation (2.120) are based on the Axilrod-Teller-Muto [219, 220] model and given by

$$E^{\text{D3,(3)}} = -\frac{1}{6} \sum_{A \neq B \neq C} \frac{C_9^{ABC} (3 \cos(\theta_a) \cos(\theta_b) \cos(\theta_c) + 1)}{(r_{AB}r_{BC}r_{CA})^3} f_{d,(3)}(\bar{r}_{ABC}), \quad (2.125)$$

where  $\theta_a, \theta_b$ , and  $\theta_c$  denote the internal angles of the triangle formed by  $r_{AB}, r_{BC}$ , and  $r_{CA}$ . The damping function  $f_{d,(3)}$  is similar to the one in Equation (2.124),  $\bar{r}_{ABC}$  is the geometric mean of  $r_{AB}, r_{BC}$ , and  $r_{CA}$ , and the  $C_9^{ABC}$  coefficient is approximated as  $C_9^{ABC} \approx -\sqrt{C_6^{AB} C_6^{BC} C_6^{CA}}$ .

Within the description of the parameter-free Tkatchenko-Scheffler vdW<sup>TS</sup> model, only pairwise dispersion energy corrections and dipole-dipole contributions are considered, meaning the specific expression of Equations (2.120) and (2.121) is given by

$$E^{\text{vdW}^{\text{TS}}} = -\frac{1}{2} \sum_{A \neq B} \frac{C_6^{AB}}{r_{AB}^6} f_{\text{damp}}(r_{AB}, r_A^0, r_B^0), \quad (2.126)$$

where  $r_{AB}$  again denotes the distance between atoms  $A$  and  $B$ , and  $r_A^0$  and  $r_B^0$  are the vdW radii. The Fermi-type damping function  $f_{\text{damp}}$  is given by [221]

$$f_{\text{damp}}(r_{AB}, r_A^0, r_B^0) = \frac{1}{1 + \exp\left[-d \left(\frac{r_{AB}}{s_R(r_A^0 + r_B^0)} - 1\right)\right]}, \quad (2.127)$$

where the free parameter  $d$  has been set to  $d = 20$  and the free empirical scaling coefficient  $s_R$  regulates the onset of the vdW correction for a specific DFA xc functional and is obtained by fitting to the S22 database of Jurečka *et al.* [222]. The atomic vdW radius  $r_A^0$  of atom  $A$  in

### 2.3. Description of the Potential Energy Surface (PES)

Equation (2.127) is expressed as

$$r_A^0 = \left( \frac{V_A}{V_A^{\text{free}}} \right)^{1/3} r_A^{0,\text{free}}, \quad (2.128)$$

where  $r_A^{0,\text{free}}$  is defined for any atom  $A$  as the radius that corresponds to the electron density contour value determined for the noble gas on the same period using its vdW radius by Bondi [39].  $\frac{V_A}{V_A^{\text{free}}}$  thereby denotes the effective atomic volume referenced to the free atom *in vacuo*, and is defined using the atomic Hirshfeld partitioning scheme [223–225]:

$$\frac{V_A}{V_A^{\text{free}}} = \frac{\int r^3 w_A(\vec{r}) \rho(\vec{r}) d\vec{r}}{\int r^3 \rho_A^{\text{free}}(\vec{r}) d\vec{r}}, \quad (2.129)$$

where  $r^3$  denotes the cube of the distance from the nucleus of atom  $A$ ,  $\rho(\vec{r})$  denotes the total electron density, and  $\rho_A^{\text{free}}(\vec{r})$  is the electron density of the free atom  $A$ . The Hirshfeld atomic partitioning weight  $w_A(\vec{r})$  is given by

$$w_A(\vec{r}) = \frac{\rho_A^{\text{free}}(\vec{r})}{\sum_B \rho_B^{\text{free}}(\vec{r})}. \quad (2.130)$$

Based on the Casimir-Polder formalism [218], a rewriting of the London formula [226] yields an expression for the heteroatomic dispersion coefficients  $C_6^{AB}$  in Equation (2.126) in terms of the homoatomic dispersion coefficients  $C_6^{AA}$  and  $C_6^{BB}$ :

$$C_6^{AB} = \frac{2C_6^{AA}C_6^{BB}}{\left[ \frac{\alpha_B^0}{\alpha_A^0} C_6^{AA} + \frac{\alpha_A^0}{\alpha_B^0} C_6^{BB} \right]}. \quad (2.131)$$

$\alpha_A^0$  and  $\alpha_B^0$  thereby denote the atomic static polarizabilities of atoms  $A$  and  $B$ . Following the same approach as in Equations (2.128) and (2.129), the atomic static polarizability  $\alpha_A^0$  of any atom  $A$  is scaled in reference to its free-atom reference value  $\alpha_A^{0,\text{free}}$  taken from the database of Chu and Dalgarno [227]:

$$\alpha_A^0 = \frac{V_A}{V_A^{\text{free}}} \alpha_A^{0,\text{free}}, \quad (2.132)$$

where  $\frac{V_A}{V_A^{\text{free}}}$  is defined in Equation (2.129). The similar expression for the homoatomic dispersion coefficients  $C_6^{AA}$  is given by

$$C_6^{AA} = \left( \frac{V_A}{V_A^{\text{free}}} \right)^2 C_6^{AA,\text{free}}, \quad (2.133)$$

where the free-atom reference values  $C_6^{AA,\text{free}}$  are again taken from the database of Chu and Dalgarno [227].

In contrast to the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> scheme that ignores the intrinsic many-body nature of correlation effects, the many-body dispersion scheme labeled MBD [228] (and sometimes also labeled MBD\* or MBD@rsSCS) combines the TS scheme with the self-consistent screening (SCS) equation of classical electrodynamics [229]. In addition, a range-separation (rs) Coulomb interaction technique is applied, separating correlation into a short-range and a long-range contribution. Short-range correlation is accounted for using a (semi-)local or hybrid xc functional, while long-range contributions are accounted for using a random-phase approximation model based on a system of localized quantum harmonic oscillators coupled in the dipole approximation [230]. Only one empirical parameter is required for the range separation step that is obtained by fitting to accurate quantum chemistry benchmark data. Computational costs are still very little in comparison to the self-consistent Kohn-Sham calculation.

*A posteriori* van der Waals correction schemes are not exclusive to DFT but may also be applied in a similar fashion for semi-empirical quantum chemistry methods in order to provide a more accurate description. As an example, the PM6 approximation, described in Subsection 2.3.4, can easily be enhanced using the unmodified D3 method by Grimme and coworkers, but using PM6-specific empirical variables [231], resulting in a method labeled PM6-D3. Based on that, Řezáč and Hobza added an additional hydrogen-bonding correction [231] that has been parameterized on the S66 benchmark data set [232]. The resulting method that also includes a correction for the underestimation of non-covalently bound atoms [233] is accordingly labeled PM6-D3H4.

### 2.3.7 Electron Correlation and Møller-Plesset Perturbation Theory

Apart from empirical methods like force fields described in Subsection 2.3.1, semi-empirical quantum chemistry methods described in Subsection 2.3.4, and the various density-functional approximations described in Subsection 2.3.5, there exists another important group of quantum chemistry methods that are wavefunction-based and developed to build on the Hartree-Fock method described in Subsection 2.3.3. Such methods are accordingly labeled post-Hartree-Fock methods [234, 235]. Their main purpose is to accurately describe the amount of electron correlation that the Hartree-Fock theory fails to adequately represent, a direct consequence of the mean-field approach in Hartree-Fock theory as the assumption that the non-interacting electrons are moving in an average potential of the other electrons neglects the tendency of “avoiding” each other more than the Hartree-Fock theory would suggest [236]. Without taking into account electronic correlation, the Hartree-Fock method fails to describe – even qualitatively – the physics of strongly correlated electrons which are *e.g.* essential in hydrogen-bonded and dispersive systems involving biomolecules [237]. The electronic ground-state correlation energy  $E_e^{\text{corr}}$  can formally be expressed as

$$E_e^{\text{corr}} = E_e - E_e^{\text{HF}}, \quad (2.134)$$

### 2.3. Description of the Potential Energy Surface (PES)

---

where  $E_e$  denotes the formally exact electronic energy in Born-Oppenheimer approximation (see Equation (2.37)) and  $E_e^{\text{HF}}$  is the Hartree-Fock energy given in Equation (2.47). There exist a variety of methods that aim to accurately determine  $E_e^{\text{corr}}$ . One popular and straightforward approach consists of treating the electronic correlation as a “small” perturbation to the Hartree-Fock wavefunction. This approach is fundamentally based on the many-body perturbation theory, or also denoted Rayleigh-Schrödinger perturbation theory [238, 239], and was later specifically formulated for Hartree-Fock wavefunctions, resulting in the Møller-Plesset perturbation theory [240]. Within the description of many-body perturbation theory, the “true” electronic Hamiltonian  $\hat{H}_e$  from Equation (2.37) is expressed as a sum of an “unperturbed” Hamiltonian  $\hat{H}_0$  and a “small” perturbation potential  $\hat{V}$ , *i.e.*

$$\hat{H}_e = \hat{H}_0 + \underbrace{\lambda \hat{V}}_{= \hat{H}'}, \quad (2.135)$$

where  $\lambda$  is a perturbation parameter with  $0 \leq \lambda \leq 1$ . The solution to the electronic Schrödinger equation of the “unperturbed” system

$$\hat{H}_0 \Psi_i^{(0)} = E_i^{(0)} \Psi_i^{(0)}, \quad (2.136)$$

is known, *i.e.* the  $\Psi_i^{(0)}$  are the corresponding obtained eigenfunctions and the  $E_i^{(0)}$  their associated energies. In order to find a solution to the Schrödinger equation of the “perturbed” system

$$\hat{H}_0 \Psi_i = E_i \Psi_i, \quad (2.137)$$

the eigenfunctions  $\Psi_i$  and their associated energies  $E_i$  are expressed in powers of  $\lambda$ , *i.e.*

$$\begin{aligned} \Psi_i &= \Psi_i^{(0)} + \lambda \Psi_i^{(1)} + \lambda^2 \Psi_i^{(2)} + \dots = \sum_{n=0} \lambda^n \Psi_i^{(n)}, \\ E_i &= E_i^{(0)} + \lambda E_i^{(1)} + \lambda^2 E_i^{(2)} + \dots = \sum_{n=0} \lambda^n E_i^{(n)}. \end{aligned} \quad (2.138)$$

Assuming intermediate normalization that can always be constructed, *i.e.*

$$\begin{aligned} \int \Psi_i^{*(0)} \Psi_i^{(n)} d\vec{r}^N &= \delta_{n0}, \\ \int \Psi_i^{*(0)} \Psi_i d\vec{r}^N &= 1, \end{aligned} \quad (2.139)$$

one yields equations for the first-order energy correction  $E_i^{(1)}$ , second-order energy correction  $E_i^{(2)}$ , etc. [236,241]:

$$\begin{aligned}
 E_i^{(0)} &= \int \Psi_i^{(0)} \hat{H}_0 \Psi_i^{(0)} d\vec{r}^N, \\
 E_i^{(1)} &= \int \Psi_i^{(0)} \hat{V} \Psi_i^{(0)} d\vec{r}^N, \\
 E_i^{(2)} &= \int \Psi_i^{(0)} \hat{V} \Psi_i^{(1)} d\vec{r}^N, \\
 E_i^{(3)} &= \int \Psi_i^{(0)} \hat{V} \Psi_i^{(2)} d\vec{r}^N, \\
 &\vdots \\
 E_i^{(n)} &= \int \Psi_i^{(0)} \hat{V} \Psi_i^{(n-1)} d\vec{r}^N.
 \end{aligned} \tag{2.140}$$

In other words, in order to find the  $n$ -th-order energy correction  $E_i^{(n)}$ , the  $(n-1)$ -th-order wavefunction correction  $\Psi_i^{(n-1)}$  is required. Within the description of Møller-Plesset perturbation theory, the “unperturbed” Hamiltonian is given by

$$\hat{H}_0 = \sum_{k=1}^N \hat{F}_k, \tag{2.141}$$

where the  $\hat{F}_k$  denote the one-electron Fock operators defined in Equation (2.55). The ground-state Hartree-Fock wavefunction  $\Psi_0^{(0)}$  is a Slater determinant (see Equation (2.46)) and an eigenfunction of  $\hat{H}_0$ . Its corresponding ground-state energy  $E_0^{(0)}$  is just the sum of orbital energies  $\epsilon_k$  (see Equation (2.60)) for the  $N$  occupied orbitals [30], *i.e.*

$$E_0^{(0)} = \sum_{k=1}^{\text{occupied}} \epsilon_k. \tag{2.142}$$

However, the ground-state Hartree-Fock wavefunction  $\Psi_0^{(0)}$  is just one of the eigenfunctions  $\Psi_i^{(0)}$  of  $\hat{H}_0$ . The system has not only  $N$  occupied spin-orbitals but also virtual ones. Since the Fock operators  $\hat{F}_k$  (and thus  $\hat{H}_0$ ) are hermitian, a complete set of eigenfunctions of  $\hat{H}_0$  exists, namely all possible spin-orbital functions that can be made up of all possible products of any  $N$  of the occupied and virtual spin-orbitals. Obviously, these eigenfunctions need to be expressed as antisymmetric Slater determinants (see Equation (2.46)). The perturbation  $\hat{H}'$  of the system is given by

$$\begin{aligned}
 \hat{H}' &= \hat{H}_e - \hat{H}_0 \\
 &= \hat{H}_e - \sum_{k=1}^N \hat{F}_k \\
 &= \frac{1}{2} \sum_{k_1 \neq k_2=1}^N \frac{1}{|\vec{r}_{k_1} - \vec{r}_{k_2}|} + \sum_{l=1}^N [\hat{j}_l - \hat{k}_l],
 \end{aligned} \tag{2.143}$$

where the one-electron Coulomb operator  $\hat{j}_l$  and the one-electron exchange operator  $\hat{k}_l$  are given in Equations (2.56) and (2.57), respectively. Evaluating to first-order in Møller-Plesset

## 2.3. Description of the Potential Energy Surface (PES)

perturbation theory then yields for the ground state [30]

$$E_0^{(0)} + E_0^{(1)} = E_e^{\text{HF}}, \quad (2.144)$$

where  $E_e^{\text{HF}}$  denotes the Hartree-Fock energy given in Equation (2.47). In other words, ground-state correlation corrections beyond Hartree-Fock theory as motivated in Equation (2.134) require second-order Møller-Plesset perturbation theory, commonly abbreviated MP2. Denoting occupied spin-orbitals obtained from Hartree-Fock theory with  $\psi_i, \psi_j$ , *etc.*, unoccupied (virtual) spin-orbitals with  $\psi_\alpha, \psi_\beta$ , *etc.*, and  $\epsilon_i, \epsilon_j, \epsilon_\alpha, \epsilon_\beta$ , *etc.* are the corresponding orbital energies, the second-order energy correction for the ground state is then given by

$$E_0^{(2)} = \frac{1}{4} \sum_{i,j}^{\text{occupied}} \sum_{a,b}^{\text{virtual}} \frac{\left| \int \int \frac{\psi_i^*(\vec{r}) \psi_j^*(\vec{r}') [\psi_\alpha(\vec{r}) \psi_\beta(\vec{r}') - \psi_\beta(\vec{r}) \psi_\alpha(\vec{r}')] }{|\vec{r} - \vec{r}'|} d\vec{r} d\vec{r}' \right|^2}{\epsilon_i + \epsilon_j - \epsilon_\alpha - \epsilon_\beta}. \quad (2.145)$$

MP2 calculations formally scale with  $\mathcal{O}(N^5)$  [123]. One appealing feature of MP2 is the obvious inclusion of many-body correlation effects which has also been made use of in Subsection 2.3.5 where an MP2-like contribution in Equation (2.115) has been included in the description of double hybrid xc functionals. However, MP2 generally tends to overestimate the correlation interaction energy in clusters [242] and fails to describe semiconductor or metallic systems due to the small or vanishing band gaps resulting in a break down of the perturbation approach. Increased computational costs for laying out higher-order Møller-Plesset perturbation theory calculations, *i.e.* MP3, MP4, *etc.*, are generally not justified due to the tendency to not improve, or even diverge, the energetic description of the system [243].

### 2.3.8 Configuration Interaction and Coupled-Cluster Theory

Within the description of Møller-Plesset perturbation theory in the last subsection, the formally complete set of eigenfunctions of the system has been introduced, meaning all possible spin-orbital functions that can be made up of all possible products of any  $N$  of the occupied and virtual spin-orbitals. The wavefunction  $\Psi_e^{\text{HF}}$  obtained from Hartree-Fock theory is a Slater determinant (see Equation (2.46)) made up of a product of  $N$  occupied spin-orbitals  $\psi_i, i = 1, \dots, N$ . By “replacing” occupied spin-orbitals with unoccupied (virtual) spin-orbitals in the Slater determinant, a whole series of Slater determinants may be created. Such Slater determinants that have one occupied spin-orbital replaced with a virtual one are denoted “singly excited”, or just “Singles” Slater determinants. Such Slater determinants that have two occupied spin-orbitals replaced with virtual ones are denoted “doubly excited”, or just “Doubles” Slater determinants. Such Slater determinants that have three occupied spin-orbitals replaced with virtual ones are denoted “triply excited”, or just “Triples” Slater determinants, *etc.* Including all possibilities of creating “excited” Slater determinants as well as ensuring the CBS limit as laid out in Subsection 2.3.3 means recovering the complete electronic correlation and formally solving the Schrödinger equation. In other words, the more “excited” Slater determinants are included and the more “complete” the basis set, the more accurate the results will

be. The method of configuration interaction (CI) consists of following the same variational principle as in Hartree-Fock theory (see Subsection 2.3.3) but using the wavefunction *Ansatz*

$$\Psi_e^{\text{CI}} = c_0 \Psi_e^{\text{HF}} + \sum_i^{\text{occupied virtual}} \sum_{\alpha} c_i^{\alpha} \Psi_i^{\alpha} + \sum_{i,j}^{\text{occupied virtual}} \sum_{\alpha,\beta} c_{i,j}^{\alpha,\beta} \Psi_{i,j}^{\alpha,\beta} + \dots, \quad (2.146)$$

where  $\Psi_e^{\text{HF}}$  denotes the Slater determinant wavefunction from Hartree-Fock theory,  $\Psi_i^{\alpha}$  denotes the Singles Slater determinant where the occupied spin-orbital  $i$  has been replaced by the virtual spin-orbital  $\alpha$ ,  $\Psi_{i,j}^{\alpha,\beta}$  denotes the Doubles Slater determinant where the two occupied spin-orbitals  $i$  and  $j$  have been replaced by the virtual spin-orbitals  $\alpha$  and  $\beta$ , *etc.* Applying the variational principle of Equation (2.41) means minimizing the energy

$$E[\Psi_e^{\text{CI}}] = \frac{\langle \Psi_e^{\text{CI}} | \hat{H}_e | \Psi_e^{\text{CI}} \rangle}{\langle \Psi_e^{\text{CI}} | \Psi_e^{\text{CI}} \rangle} \quad (2.147)$$

by varying the linear coefficients  $c_{\dots}$  in Equation (2.146) under the constraint that  $\Psi_e^{\text{CI}}$  is normalized which can always be guaranteed by making use of the method of Lagrangian undetermined multipliers, as laid out in Subsection 2.3.3. Following this straight-forward approach then results in a general matrix eigenvalue problem that can formally be solved by diagonalizing the so-called CI matrix [30, 241]. However, in practice the number of possible Slater determinants becomes very large even for the most modest of systems. In the limit of a complete basis set, the computational costs formally scale exponentially with system size [236]. A second very important drawback consists of the method not being size-consistent: When truncating the expansion in Equation (2.146), this results in the formal consequence that the energy of  $N$  non-interacting atoms does not equal to  $N$  times the energy of a single atom, thus making CI a progressively less accurate method with increasing system size [244].

A different approach with a similar *Ansatz* is given by the coupled-cluster (CC) method [245–247] that is not based on the variational principle but guarantees size-consistency [126]. Defining the excitation operator  $\hat{T}$  as

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \dots + \hat{T}_N, \quad (2.148)$$

where  $N$  denotes the number of electrons, and the  $n$ -th excitation operator  $T_n$  acts on the wavefunction Slater determinant  $\Psi_e^{\text{HF}}$  obtained from Hartree-Fock theory by creating all possible  $n$ -times excited Slater determinants, *i.e.*

$$\begin{aligned} \hat{T}_1 \Psi_e^{\text{HF}} &= \sum_i^{\text{occupied virtual}} \sum_{\alpha} t_i^{\alpha} \Psi_i^{\alpha}, \\ \hat{T}_2 \Psi_e^{\text{HF}} &= \sum_{i < j}^{\text{occupied virtual}} \sum_{\alpha < \beta} t_{i,j}^{\alpha,\beta} \Psi_{i,j}^{\alpha,\beta}, \\ &\vdots \end{aligned} \quad (2.149)$$



### 2.3. Description of the Potential Energy Surface (PES)

where the to-be-determined linear coefficients  $t_{i,j}$  are commonly called “excitation” amplitudes, Equation (2.146) can be expressed as

$$\begin{aligned}\Psi_e^{\text{CI}} &= (\hat{1} + \hat{T})\Psi_e^{\text{HF}} \\ &= (\hat{1} + \hat{T}_1 + \hat{T}_2 + \dots)\Psi_e^{\text{HF}}.\end{aligned}\quad (2.150)$$

Within the description of the CC method however, the wavefunction *Ansatz* is given by

$$\begin{aligned}\Psi_e^{\text{CC}} &= e^{\hat{T}}\Psi_e^{\text{HF}} \\ &= \left(\hat{1} + \hat{T} + \frac{1}{2}\hat{T}^2 + \frac{1}{3!}\hat{T}^3 + \dots\right)\Psi_e^{\text{HF}} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\hat{T}^k\Psi_e^{\text{HF}}.\end{aligned}\quad (2.151)$$

Using this *Ansatz* and assuming orthonormality of  $\Psi_e^{\text{CC}}$ , the total energy of the system is then estimated as [126]

$$\begin{aligned}E[\Psi_e^{\text{CC}}] = E_e^{\text{CC}} &= \langle \Psi_e^{\text{CC}} | \hat{H}_e | \Psi_e^{\text{CC}} \rangle = \langle \Psi_e^{\text{HF}} | e^{-\hat{T}} \hat{H}_e e^{\hat{T}} | \Psi_e^{\text{HF}} \rangle \\ &= \langle \Psi_e^{\text{HF}} | \hat{H}_e e^{\hat{T}} | \Psi_e^{\text{HF}} \rangle.\end{aligned}\quad (2.152)$$

Denoting occupied spin-orbitals obtained from Hartree-Fock theory with  $\psi_i, \psi_j$ , etc. and unoccupied (virtual) spin-orbitals with  $\psi_\alpha, \psi_\beta$ , etc., expanding Equation (2.152) then yields [126]

$$\begin{aligned}E_e^{\text{CC}} = E_e^{\text{HF}} &+ \sum_{i < j}^{\text{occupied}} \sum_{\alpha < \beta}^{\text{virtual}} \left( t_{i,j}^{\alpha,\beta} + t_i^\alpha t_j^\beta - t_i^\beta t_j^\alpha \right) \left( \int \int \frac{\psi_i^*(\vec{r})\psi_j^*(\vec{r}')\psi_\alpha(\vec{r})\psi_\beta(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r} d\vec{r}' \right. \\ &\quad \left. - \int \int \frac{\psi_i^*(\vec{r})\psi_j^*(\vec{r}')\psi_\beta(\vec{r})\psi_\alpha(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r} d\vec{r}' \right),\end{aligned}\quad (2.153)$$

meaning the CC correlation energy is completely determined by the Singles amplitudes  $t_i^\alpha$ , Doubles amplitudes  $t_{i,j}^{\alpha,\beta}$ , and the two-electron integrals using spin-orbitals from Hartree-Fock theory. The Singles and Doubles amplitudes are determined by expanding the entities

$$\begin{aligned}0 &= \langle \Psi_i^\alpha | e^{-\hat{T}} \hat{H}_e e^{\hat{T}} | \Psi_e^{\text{HF}} \rangle, \\ 0 &= \langle \Psi_{i,j}^{\alpha,\beta} | e^{-\hat{T}} \hat{H}_e e^{\hat{T}} | \Psi_e^{\text{HF}} \rangle, \\ 0 &= \langle \Psi_{i,j,k}^{\alpha,\beta,\gamma} | e^{-\hat{T}} \hat{H}_e e^{\hat{T}} | \Psi_e^{\text{HF}} \rangle, \\ &\vdots\end{aligned}\quad (2.154)$$

leading to a set of coupled non-linear equations for the Singles and Doubles amplitudes that are required to be solved iteratively [30]. While the approach is formally exact, in practice the excitation operator  $\hat{T}$  needs to be truncated at some excitation level in order to make the calculation even feasible. Using the Singles (S) and Doubles (D) excitation levels, *i.e.*

$\hat{T} = \hat{T}_1 + \hat{T}_2$ , results in the CCSD model [248] that formally scales with  $\mathcal{O}(N^6)$  [123]. The operator  $e^{\hat{T}}$  in Equation (2.151) is then given by

$$e^{\hat{T}} = e^{\hat{T}_1 + \hat{T}_2} = \hat{1} + \hat{T}_1 + (\hat{T}_2 + \frac{1}{2} \hat{T}_1^2) + (\hat{T}_1 \hat{T}_2 + \frac{1}{6} \hat{T}_1^3) + (\frac{1}{2} \hat{T}_2^2 + \frac{1}{2} \hat{T}_1^2 \hat{T}_2 + \frac{1}{24} \hat{T}_1^4) + \dots \quad (2.155)$$

In contrast to Equation (2.150) for the CI *Ansatz* wavefunction, Equation (2.155) also contains higher excitation terms beyond the truncation level that are made up of so-called “disconnected” excitations, thus effectively making the CC method size-consistent. For example, although “connected” quadruple excitations ( $\hat{T}_4$ ) are not present in Equation (2.155) due to the truncation, quadruple excitations can still be made up of two “disconnected” double excitations ( $\hat{T}_2^2$ ), four “disconnected” single excitations ( $\hat{T}_1^4$ ), or a mixture of “disconnected” single and double excitations ( $\hat{T}_1^2 \hat{T}_2$ ). Using the Singles (S), Doubles (D), and Triples (T) excitation levels, *i.e.*  $\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3$ , results in the CCSDT model [249] that formally scales with  $\mathcal{O}(N^8)$ , resulting in very demanding computational costs even for modest systems. When treating Triples (T) excitation levels using Møller-Plesset perturbation theory, the resulting CCSD(T) [250] approach then formally scales with  $\mathcal{O}(N^7)$  and commonly provides excellent accuracy for non-covalent complexes [251, 252]. Hence, CCSD(T) is sometimes referred to as the “gold standard of quantum chemistry”.

Still, in recent years considerable effort has been made in order to reduce computational costs of CCSD(T) calculations without having to sacrifice accuracy. The domain-based local pair natural orbital (DLPNO)-CCSD(T) [253, 254] approximation aims to fully exploit locality of the electron correlation and shows a near-linear scaling behavior with system size  $N$ . In short, the correlation energy of the system is expressed as a sum over the correlation energies of pairs ( $ij$ ) of electrons. In case the corresponding canonical orbitals ( $i$ ) and ( $j$ ) in the Slater determinant are localized, the associated pair correlation energy  $\epsilon_{ij}$  falls off quickly and essentially non-contributing separated electron pairs are being characterized using a fast-to-compute multipole estimate screening mechanism. Furthermore, “weakly-contributing” electron pairs that have not been screened out but lay beyond some cut-off value are removed from being treated exactly. This cut-off is evaluated from estimating the pair correlation energy from MP2 calculations. Since the conventional MP2 method formally scales with  $\mathcal{O}(N^5)$  (see Subsection 2.3.7), a “semi-local” approximation is applied based on the local MP2 method with use of density fitting [255]. The virtual space is thereby expanded in so-called projected atomic orbitals (PAOs) [256] that are locally associated with the “parent” atomic domain. The “weakly-contributing” electron pairs beyond the cut-off are treated using the “semi-local” MP2 approximation that formally scales linearly in computational costs, thereby effectively making up a large amount of error introduced due to the truncation scheme. When laying out the evaluations for the “strongly-contributing” electron pairs, one generally wishes to limit excitation amplitudes to only those associated with occupied orbitals ( $i$ ) and ( $j$ ) and the local domain associated with the electron pair ( $ij$ ). This is achieved by making use of approximate natural orbitals of a given electron pair, so-called pair natural orbitals (PNOs) [257–259], that are created from an approximate pair density matrix evaluated using the previously mentioned

## 2.3. Description of the Potential Energy Surface (PES)

---

“semi-local” MP2 approximation. PNOs with a low occupation number beyond a cut-off value are neglected. While PNOs are locally associated with the occupied orbitals ( $i$ ) and ( $j$ ), they are also by construction “delocally expanded” into the virtual space according to the correlation of the electron pair ( $ij$ ). The PNOs are then expanded in terms of the local PAOs which allows for efficiently restricting the evaluation of the excitation amplitudes to newly defined local domains associated with the electron pair ( $ij$ ), which is controlled by a third cut-off parameter. Finally, within the description of treating the Triples (T) excitation levels in a perturbative manner, so-called “triples-natural orbitals” (TNOs) are created for each considered electron triple ( $ijk$ ). By construction, the TNOs thereby span the significant PNO subspace of the three electron pairs ( $ij$ ), ( $jk$ ), and ( $ik$ ), for which at least one of these must correspond to a “weakly-contributing” electron pair in order for the electron triple ( $ijk$ ) to be considered [254].

### 2.3.9 Computer Simulations and Practical Considerations

Computer simulations are carried out using different software depending on the theoretical method being applied. All FF calculations are done using the TINKER molecular modeling package [260]. The applied version 7.1.2 contains out-of-the-box parameter files available for all force fields described in Subsection 2.3.1. Single-point energy evaluations for the semi-empirical quantum-chemistry methods mentioned in Subsection 2.3.4 are carried out using the MOPAC2016 [261] semi-empirical quantum chemistry program. In contrast to the other methods for which energy calculations refer to total energies on the potential-energy surface, semi-empirical energy evaluations yield heats of formation as the respective semi-empirical methods are parameterized on experimental heats of formation [262]. The heat of formation is thereby defined as the sum of the electronic energy, the nuclear-nuclear repulsion energy, the ionization energy for the valence electrons, the total heat of atomization of all the atoms in the system, and – if available – the energy from hydrogen bonds and dispersion correction [263].

Evaluations that involve different DFAs (see Subsection 2.3.5) are almost entirely done using the all-electron/full-potential electronic structure code package FHI-aims [264, 182, 265]. The algorithms for *ab initio* molecular simulations within FHI-aims are based on basis sets that are numerically tabulated and centered at each atom composing the system being studied, hence them being labeled as numerically tabulated atom-centered orbitals (NAOs). The basis functions are thereby organized in so-called tiers, *i.e.* levels of basis function groups that arose from a basis optimization procedure such that their ordering reflects the amount of improvement on the element-dependent absolute convergence levels achieved. For example, for light elements, *e.g.* carbon or oxygen, tier 1 basis sets guarantee accurate geometry (pre-)relaxations, and tier 2 basis sets are required to guarantee meV-converged<sup>1</sup> energy differences, while for heavier elements such convergence levels are often already yielded using tier 1 basis sets. Because of this, a hierarchy of predefined settings called `light`, `tight`, and `really-tight` is provided for all elements. While `light` settings are generally used for geometry (pre-)relaxations, `tight` settings are commonly used for production runs as they

---

<sup>1</sup>1 meV  $\approx$  0.023 kcal/mol

usually already guarantee meV-converged energy differences. Access to all elements from light to heavy is ensured by making use of a scalar-relativistic treatment in applying the scaled zeroth order regular approximation (ZORA) [264, 266] scheme. Finally, FHI-aims is open to be used with different basis set families. For example, numerically tabulated atom-centered orbital  $n$ -zeta basis sets with valence-correlation consistency, labeled NAO-VCC-nZ [267], have been specifically constructed to be used for methods that invoke the continuum of unoccupied orbitals explicitly, *e.g.* MP2 (see Subsection 2.3.7) or double hybrid DFAs (see Subsection 2.3.5). Constructed analogous to Dunning's correlation-consistent polarized valence-only basis sets (cc-pVnZ) [268, 269], these basis sets utilize the more flexible shape of NAOs, hence both the behavior near the nucleus as well as that for the tails of orbitals far away from atoms is intended to be much more physical. In particular, Zhang *et al.* showed that the double hybrid DFA XYG3 provides best results in combination with the triple-zeta NAO-VCC-3Z basis set [270].

Calculations for wavefunction-based methods, *i.e.* coupled-cluster calculations (see Subsection 2.3.8) and MP2 (see Subsection 2.3.7), are carried out with the electronic structure program package ORCA [271] using Ahlrichs' def2 [200] basis set family. Because heavy elements like  $\text{Zn}^{2+}$  require a relativistic treatment, the ZORA scheme is implemented in ORCA in an approximate way [272, 273]. As the scalar relativistic treatment requires flexible basis sets, this in turn means that ORCA automatically provides relativistically recontracted versions [274] of Ahlrichs' def2 basis set family, labeled ZORA-def2. In practice however, wavefunction-based methods come with a severe limiting feature concerning their accuracy, namely their slow convergence of correlation energy calculations to the complete basis set (CBS) limit [125, 275]. In order to account for that, extrapolation schemes for systematic convergent basis set families, *e.g.* basis set families by Dunning *et al.* or Ahlrichs *et al.* (def2), may be applied. For example, Hartree-Fock energies may be extrapolated using a form proposed by Karton and Martin [276]:

$$E_n^{\text{HF}} = E_{\text{CBS}}^{\text{HF}} + A e^{-\alpha\sqrt{n}}, \quad (2.156)$$

with  $A$ ,  $\alpha$ , and the CBS-extrapolated energy  $E_{\text{CBS}}^{\text{HF}}$  being parameters to be determined from a least-squares fitting algorithm. The cardinal number  $n$  thereby denotes the respective basis set hierarchy, *i.e.*  $n = 2$  for double-zeta basis sets,  $n = 3$  for triple-zeta basis sets, *etc.* A similar extrapolation scheme may also be laid out for the correlation energies following the form proposed by Truhlar [275]:

$$E_n^{\text{corr}} = E_{\text{CBS}}^{\text{corr}} + B n^{-\beta}, \quad (2.157)$$

again with  $B$ ,  $\beta$ , and the CBS-extrapolated energy  $E_{\text{CBS}}^{\text{corr}}$  being parameters to be determined from a least-squares fitting algorithm as before. Assuming  $\beta = 3$  yields an effective two-point extrapolation scheme as originally proposed by Halkier *et al.* [277].

Another related issue for wavefunction-based methods that comes with slow-converging correlation contributions due to the usage of finite basis sets is given by the basis set superposition error (BSSE) [203, 204]: When atoms are bonded together in a molecule, the usage of finite basis sets then leads to artificially more stable energies because of their availability to

## 2.4. Conformational Sampling and Basin-Hopping

overlapping basis functions belonging to other nearby components beyond their own basis functions. When for example comparing relative energies between different conformers, this may lead to large energetic discrepancies depending on the specific structures in place. To account for that and prior to performing CBS extrapolation as described above, one may subject the Hartree-Fock and correlation energies to a counterpoise correction as proposed by Boys and Bernardi [278]: Assuming rigid conformers, the BSSE between two components (labeled Comp1 and Comp2) is estimated as

$$\begin{aligned} E_{\text{BSSE}} &= E_{\text{BSSE}}(\text{Comp1}) + E_{\text{BSSE}}(\text{Comp2}), \\ \text{with } E_{\text{BSSE}}(\text{Comp1}) &= E^{\text{Comp1+Comp2}}(\text{Comp1}) - E^{\text{Comp1}}(\text{Comp1}), \\ \text{and } E_{\text{BSSE}}(\text{Comp2}) &= E^{\text{Comp1+Comp2}}(\text{Comp2}) - E^{\text{Comp2}}(\text{Comp2}), \end{aligned} \quad (2.158)$$

where  $E^{\text{Comp1+Comp2}}(\text{Comp1})$  represents the energy of component 1 evaluated in the union of the basis functions associated with component 1 and component 2,  $E^{\text{Comp1}}(\text{Comp1})$  represents the energy of component 1 evaluated in the basis functions associated with component 1, *etc.* The individual BSSEs are then to be subtracted from the Hartree-Fock and correlation energy, respectively.

## 2.4 Conformational Sampling and Basin-Hopping

Subsections 2.3.1 through 2.3.8 provided an overview on how to evaluate single-point energies on the PES using vastly different theoretical models and levels of theory. As motivated in Section 2.2, the description of the free energy landscape also requires an accurate sampling of the PES – at least near the global minimum – from which following quantities like free energies, vibrational modes, *etc.* are derived. While the latter will be laid out in Section 2.5, this subsection focusses on sampling approaches for systematically surveying the PES. Ultimately, the interest lies in characterizing the global minimum region (and eventual regions of “kinetic traps”) as this is the region where the peptide is assumed its native structure with minimal free energy and entropy, following Anfinsen’s “thermodynamic hypothesis” as well as the “folding funnel hypothesis” explained in Section 2.2. Within this scenario, being able to predict the peptide’s three-dimensional structure of the native state given only the amino acid sequence then would allow for a rigorous comparison with experiment, an important factor in peptide structure elucidation that will be discussed in detail in Chapter 3 of this work. In any case, finding the global minimum within the context of peptide structure prediction is a global optimization problem for which a large variety of approaches have been suggested [69, 279, 280]. Examples include methods based on Monte Carlo simulations, *e.g.* simulated annealing [281], multi-canonical Monte Carlo sampling [282], entropic sampling [283], replica exchange Monte Carlo [284], or parallel hyperbolic sampling [285], methods based on molecular dynamics, *e.g.* replica exchange molecular dynamics [286] or accelerated molecular dynamics [287], methods following a genetic algorithm, *e.g.* implemented in FAFOOM (flexible algorithm for optimization of molecules) [288], and conformational space annealing [289]. One approach that is being extensively made use of in Chapter 3 of this work is called basin-hopping and

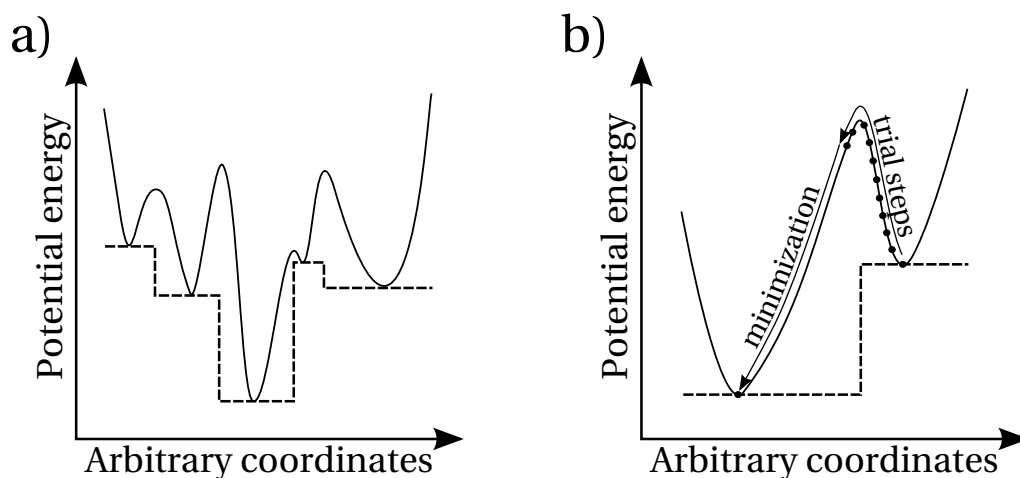


Figure 2.9 – (a) Schematic illustration of the effect of transforming the PES to plateaux of energies of local minima obtained by the basin-hopping approach. The solid and dashed lines represent the “true” potential  $V$  and the transformed potential  $\tilde{V}$ , respectively. (b) Schematic illustration of the basin-hopping approach within the “conformational scanning” algorithm.

was originally formulated as a Monte Carlo-based method [290, 291], although a simplified version being labeled “conformational scanning” [292] is being used here. With  $\vec{X}$  denoting all relevant atomic coordinates of the system, the basic idea of this approach then consists in surveying the complete PES by transforming the “true” energy potential  $V(\vec{X})$  at any given point  $\vec{X}$  into the transformed energy potential landscape  $\tilde{V}(\vec{X})$  that is just the set of nearest local minima, so-called catchment basins. In other words, a local minimization is applied at any given point  $\vec{X}$ , *i.e.*

$$\tilde{V}(\vec{X}) = \min \{V(\vec{X})\}. \quad (2.159)$$

The resulting transformed energy potential landscape  $\tilde{V}(\vec{X})$  then consists of plateaux at the energies of the local minima, as schematically illustrated in Figure 2.9(a). The associated search strategy is a simple iterative scheme of hopping between minima (basins) [292]: Starting from a random conformation, a first local minimization within an energy convergence criterion is laid out using a truncated Newton method with a preconditioned linear conjugate gradient solution of Newton’s equations [293]. This first minimum then serves as a seed for an iterative search procedure. For the corresponding conformer, the torsional space Hessian matrix is calculated and diagonalized to obtain its eigenvectors  $\vec{X}_\omega$ . Along these eigenvectors, the system is moved out of the local minimum in fixed trial steps  $k\vec{X}_\omega$  and  $-k\vec{X}_\omega$  ( $k = 1, 2, \dots, 65$ ). At each point  $k$ , the conformational energy  $E_k$  is calculated. In case of overcoming an energy barrier, a situation defined by the inequalities  $E_{k-1} > E_k$  and  $E_{k-1} > E_{k+1}$ , a local minimization is performed, as schematically illustrated in Figure 2.9(b). If the newly obtained minimum is within a predefined energy threshold and differs from previously obtained minima ensured by a simple energy comparison, it is added to the list of minima. The procedure is then repeated for every newly obtained minimum until no further minima are found, meaning

## 2.4. Conformational Sampling and Basin-Hopping

---

the whole conformational space of energy minimum plateaux has been grasped. Obviously, this approach is only feasible for rather small systems as the conformational space increases exponentially with system size [294].

## 2.5 Description of the Free Energy Surface and Comparison to Experiment

After having characterized various methods for evaluating energies on the PES in Section 2.3 using vastly different theoretical models and levels of theory, the description of the free energy surface is tackled in this section. After all, the energy on the PES can formally be derived from a fixed structure only, *e.g.* the global minimum on the PES, while deriving quantities like free energies or vibrational modes requires the description of molecules in motion, *e.g.* vibration in harmonic approximation or molecular dynamics. The general concept of free energies has already been touched on in Section 2.2. As laid out there, the work within this thesis relies on the Helmholtz free energy for which free energy contributions are accounted for from internal degrees of freedom, consisting of vibration and rotation, in addition to the potential energy on the PES. In particular, the usage of the Helmholtz free energy for free energy contributions has been motivated as being beneficial for studies of peptide systems in the gas phase, *i.e.* in isolation. Most importantly however, this quantity as well as the thereby derived vibrational spectra of peptides are of fundamental interest for comparing with experimental data obtained under certain conditions, *i.e.* for peptide ion systems in the gas phase at cold temperatures (10K), as further explained in Subsection 2.5.2.

### 2.5.1 Infrared Spectra and Free Energy Calculations in Harmonic Approximation

Before deriving expressions for the free energy contributions, it is useful to first get familiar with the description of molecular vibration in harmonic approximation as the theoretical background thereof facilitates the understanding and derivation of said expressions within the framework of quantum statistical mechanics. Within the quantum-mechanical treatment of peptide systems in the gas phase, the Born-Oppenheimer approximation was introduced in Subsection 2.3.2: Because electrons move “instantaneously” with respect to the movement of the nuclei, the many-body wavefunction  $\Psi$  of the system was approximated as (see Equation (2.30))

$$\Psi = \Psi_n \Psi_e, \quad (2.160)$$

where  $\Psi_e$  denotes the electronic wavefunction and  $\Psi_n$  denotes the nuclear wavefunction. Following this approach of separating the movements of nuclei and electrons, it is immediately clear that vibrations of molecules may be treated by taking into consideration only the movements of the  $M$  nuclei on the PES, *i.e.* described by the potential (compare to Equation (2.32))

$$V_{\text{BO}}(\vec{R}_1, \dots, \vec{R}_M) = \frac{1}{2} \sum_{k_1 \neq k_2=1}^M \frac{Z_{k_1} Z_{k_2} e^2}{|\vec{R}_{k_1} - \vec{R}_{k_2}|} + E_e(\vec{R}_1, \dots, \vec{R}_M). \quad (2.161)$$

The first term describes the nucleus-nucleus interaction of the system and the second term denotes the electronic energy for a given set of atomic coordinates  $\vec{R}_1, \dots, \vec{R}_M$ , that is obtained by solving the electronic Schrödinger equation (Equation (2.37)) for which a variety of theoretical models and levels of theory have been discussed in Subsections 2.3.1 through 2.3.8. Ultimately,



## 2.5. Description of the Free Energy Surface and Comparison to Experiment

the interest lies in the native structure of the system which is assumed in the (local) minimum of the PES. For cold peptide systems in the gas phase, vibrations of the molecule may then be proficiently described in terms of small atomic displacements from the equilibrium structure. Denoting the atomic positions in equilibrium as  $\vec{R}_1^0, \dots, \vec{R}_M^0$ , the Born-Oppenheimer potential in Equation (2.161) may then be expanded in a Taylor series, *i.e.*

$$V_{\text{BO}}(\vec{R}_1, \dots, \vec{R}_M) = V_{\text{BO}}(\vec{R}_1^0, \dots, \vec{R}_M^0) + \sum_{k=1}^M \left( \frac{\partial V_{\text{BO}}}{\partial \vec{R}_k} \right)_{\vec{R}_1^0, \dots, \vec{R}_M^0} (\vec{R}_k - \vec{R}_k^0) + \frac{1}{2} \sum_{k_1, k_2=1}^M \left( \frac{\partial^2 V_{\text{BO}}}{\partial \vec{R}_{k_1} \partial \vec{R}_{k_2}} \right)_{\vec{R}_1^0, \dots, \vec{R}_M^0} (\vec{R}_{k_1} - \vec{R}_{k_1}^0) (\vec{R}_{k_2} - \vec{R}_{k_2}^0) + \dots \quad (2.162)$$

The first term is just a constant offset of the potential and may always be set to zero for convenience. The second term vanishes for stationary points on the PES. The harmonic approximation consists of truncating the expansion at second order and may be justified for small displacements  $\vec{R}_k - \vec{R}_k^0$  ( $k = 1, \dots, M$ ), *i.e.*

$$V_{\text{BO}}(\vec{R}_1, \dots, \vec{R}_M) \approx \frac{1}{2} \sum_{k_1, k_2=1}^M \left( \frac{\partial^2 V_{\text{BO}}}{\partial \vec{R}_{k_1} \partial \vec{R}_{k_2}} \right)_{\vec{R}_1^0, \dots, \vec{R}_M^0} (\vec{R}_{k_1} - \vec{R}_{k_1}^0) (\vec{R}_{k_2} - \vec{R}_{k_2}^0). \quad (2.163)$$

Within the treatment of molecular vibrations in harmonic approximation, the concept of normal coordinates plays an important role [295]. It is advantageous to first treat the problem using classical mechanics as the yielded vibrational frequencies of the harmonic motions will give rise to quantized energy levels within the quantum mechanical description that – surprisingly or not – depend on the classical vibrational frequencies [296], as laid out below. It should be pointed out that within the description of classical mechanics, molecular vibrations are treated in terms of the coordinates of a moving system of axes that “moves and rotates with the molecule” just as if the molecule were not undergoing translation or rotating. This most importantly implies two things: (i) The problem of vibration in molecules may be treated in very good approximation independently of molecular translation and rotation, and (ii) out of the  $3M$  degrees of freedom of the system of  $M$  atoms only  $3M - 6$  are independent of each other because six conditions are required to define such a moving system of axes. A rigorous description thereof is provided in Reference [297]. Denoting the atomic masses as  $M_k$  ( $k = 1, \dots, M$ ), the mass-weighted displacement coordinates  $\vec{q}_k$  can be defined as

$$\vec{q}_k = \sqrt{M_k} (\vec{R}_k - \vec{R}_k^0), \quad (k = 1, \dots, M). \quad (2.164)$$

The kinetic energy  $T_{\text{BO}}$  of the system

$$T_{\text{BO}} = \frac{1}{2} \sum_{k=1}^M M_k \left( \frac{d(\vec{R}_k - \vec{R}_k^0)}{dt} \right)^2 \quad (2.165)$$

## Chapter 2. Theoretical and Experimental Background, Methods, and Techniques

can then be re-written as

$$T_{\text{BO}} = \frac{1}{2} \sum_{i=1}^{3M} \dot{q}_i^2, \quad (2.166)$$

where the indices  $i, j, \dots$  are now used to enumerate coordinates, and  $\dot{q}_i = \frac{dq_i}{dt}$ . Similarly, the potential energy  $V_{\text{BO}}$  in Equation (2.163) can be re-written as

$$V_{\text{BO}} = \frac{1}{2} \sum_{i,j=1}^{3M} \left( \frac{\partial^2 V}{\partial q_i \partial q_j} \right)_0 q_i q_j. \quad (2.167)$$

Newton's equations of motion

$$M_k \ddot{\vec{R}}_k = - \frac{\partial V_{\text{BO}}}{\partial \vec{R}_k}, \quad k = 1, \dots, M, \quad (2.168)$$

can be re-written as [297]

$$\ddot{q}_i = - \sum_{j=1}^{3M} \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_i \partial q_j} \right)_0 q_j, \quad i = 1, \dots, 3M. \quad (2.169)$$

This set of  $3M$  second-order linear differential equations may be solved using the *Ansatz*

$$q_i = A_i \cos(\omega t + \phi), \quad (2.170)$$

where the amplitude  $A_i$ , the frequency  $\omega$ , and the phase  $\phi$  are parameters. Substituting Equation (2.170) into Equation (2.169) yields a set of homogeneous linear algebraic equations, namely

$$\sum_{i=1}^{3M} \left[ \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_i \partial q_j} \right)_0 - \delta_{ij} \omega^2 \right] A_i = 0, \quad j = 1, \dots, 3M. \quad (2.171)$$

Only a special set of values of  $\omega^2$  gives non-trivial solutions, namely the one that satisfies the secular equation

$$|\mathbf{H} - \omega^2 \mathbf{I}| = 0, \quad (2.172)$$

where  $\mathbf{I}$  denotes the identity matrix and  $\mathbf{H}$  is the mass-weighted Hessian matrix given by

$$\mathbf{H} = \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_i \partial q_j} \right)_0 = \begin{pmatrix} \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_1 \partial q_1} \right)_0 & \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_1 \partial q_2} \right)_0 & \cdots & \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_1 \partial q_{3M}} \right)_0 \\ \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_2 \partial q_1} \right)_0 & \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_2 \partial q_2} \right)_0 & \cdots & \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_2 \partial q_{3M}} \right)_0 \\ \cdots & \cdots & \cdots & \cdots \\ \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_{3M} \partial q_1} \right)_0 & \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_{3M} \partial q_2} \right)_0 & \cdots & \left( \frac{\partial^2 V_{\text{BO}}}{\partial q_{3M} \partial q_{3M}} \right)_0 \end{pmatrix}. \quad (2.173)$$

Solving the eigenwert problem of Equation (2.172) implies finding the eigenvalues  $\omega_n^2$  and the corresponding eigenvectors  $A_{i,n}$ , where the index  $n = 1, \dots, 3M$  indicates the amplitudes' (the  $A_{i,n}$ 's) correspondence to a particular eigenvalue  $\omega_n^2$ . The set of homogeneous linear algebraic equations thereby does not determine the amplitudes  $A_{i,n}$  uniquely as they depend on the initial values of the system, *i.e.* the initial values of the coordinates  $q_i$  and velocities  $\dot{q}_i$ .

## 2.5. Description of the Free Energy Surface and Comparison to Experiment

---

Therefore defining

$$l_{i,n} = \frac{A_{i,n}}{\sqrt{\sum_i^{3M} (A_{i,n})^2}} \quad (2.174)$$

for any amplitudes  $A_{i,n}$ , allows for expressing

$$A_{i,n} = K_n l_{i,n} \quad (2.175)$$

where the  $K_n$  are constants determined by the initial values of the  $q_i$  and  $\dot{q}_i$ . The general solution then reads

$$q_i = \sum_{n=1}^{3M} K_n l_{i,n} \cos(\omega_n t + \phi_n), \quad i = 1, \dots, 3M. \quad (2.176)$$

The amplitudes  $K_n$  and the phases  $\phi_n$  are determined by the initial values of the system, *i.e.* the initial values of the coordinates  $q_i$  and velocities  $\dot{q}_i$ . Since the system has  $3M - 6$  degrees of freedom as laid out above, six of the eigenvalues  $\omega_n^2$  must be zero while the other  $3M - 6$  of the eigenvalues  $\omega_n^2$  must be non-zero and describing vibrational states of the system, meaning the eigenvectors  $A_{i,n}$  are the amplitudes of the different coordinates that oscillate with the same frequency  $\omega_n$  and phase  $\phi_n$ . The six independent modes of motion associated with the six eigenvalues  $\omega_n^2$  that are zero are corresponding to the three translations and the three rotations of the system. The corresponding sets of  $l_{i,n}$  are determined as usual from the set of homogeneous linear algebraic equations given in Equation (2.171) when setting  $\omega = 0$ .

Before treating the problem using quantum mechanics, it is very convenient to introduce normal coordinates  $Q_k$  ( $k = 1, \dots, 3M$ ) defined as

$$Q_k = \sum_{i=1}^{3M} l_{i,k} q_i, \quad k = 1, \dots, 3M, \quad (2.177)$$

using the  $l_{i,k}$  defined in Equation (2.174). Using this particular set of coordinates, the kinetic energy  $T_{\text{BO}}$  of the system from Equation (2.166) then reads

$$T_{\text{BO}} = \frac{1}{2} \sum_{k=1}^{3M} \dot{Q}_k^2, \quad (2.178)$$

while the potential energy  $V_{\text{BO}}$  from Equation (2.167) then reads

$$V_{\text{BO}} = \frac{1}{2} \sum_{k=1}^{3M} \omega_k^2 Q_k^2. \quad (2.179)$$

While the kinetic energy retains its form, the form of the potential energy simplifies tremendously as it no longer contains cross products of different coordinates. Employing the same arguments as above, one may only consider the  $3M - 6$  degrees of freedom corresponding to

molecular vibration, *i.e.*

$$T_{\text{BO}} = \frac{1}{2} \sum_{k=1}^{3M-6} \dot{Q}_k^2, \quad V_{\text{BO}} = \frac{1}{2} \sum_{k=1}^{3M-6} \omega_k^2 Q_k^2. \quad (2.180)$$

The qualitative arguments of decoupling molecular vibration, rotation, and translation carry over from classical mechanics to the quantum mechanical description, although a rigorous description is tedious [297]. The many-body wavefunction  $\Psi$  of the system in Born-Oppenheimer approximation in Equation (2.160) may then further be approximated as

$$\Psi = \Psi_e \Psi_n = \Psi_e \Psi_v \Psi_r \Psi_t, \quad (2.181)$$

where the nuclear wavefunction  $\Psi_n$  is approximated as being separable in a vibrational part  $\Psi_v$ , a rotational part  $\Psi_r$ , and a translational part  $\Psi_t$ , all while the electronic wavefunction  $\Psi_e$  does not contribute to the description as laid out in the beginning of this Subsection. Hence within this approximation, the nuclear vibrational Schrödinger equation immediately follows from Equation (2.180):

$$-\frac{1}{2} \sum_{k=1}^{3M-6} \frac{\partial \Psi_v}{\partial Q_k^2} + \frac{1}{2} \sum_{k=1}^{3M-6} \omega_k^2 Q_k^2 \Psi_v = E_v \Psi_v. \quad (2.182)$$

Having used normal coordinates greatly simplifies the problem as the solution is separable in its normal coordinates. In other words, using the *Ansatz*

$$\Psi_v = \Psi_v(Q_1) \Psi_v(Q_2) \cdots \Psi_v(Q_{3M}) \quad (2.183)$$

yields  $3M - 6$  total differential equations in one variable, the normal coordinate  $Q_k$ , *i.e.*

$$-\frac{1}{2} \frac{\partial \Psi_v(Q_k)}{\partial Q_k^2} + \frac{1}{2} \omega_k^2 Q_k^2 \Psi_v(Q_k) = E_v(k) \Psi_v(Q_k), \quad k = 1, \dots, 3M - 6, \quad (2.184)$$

where the energy  $E_v$  is given by the sum

$$E_v = E_v(1) + E_v(2) + \dots + E_v(3M - 6). \quad (2.185)$$

Equation (2.184) is the wave equation of the linear harmonic oscillator for which the solution is well-known [298, 299]. The energy of a linear harmonic oscillator is given by<sup>2</sup>

$$E_v(k) = \hbar \omega_k \left( n_k + \frac{1}{2} \right), \quad n_k = 0, 1, 2, \dots, \quad (2.186)$$

where the  $n_k$  denote quantum numbers. Note that  $\omega_k$  is the classical frequency of the system associated with the normal coordinate  $Q_k$ . The energy  $E_v$  of the set of  $3M - 6$  coupled quantum

---

<sup>2</sup>Here,  $\hbar$  has been written explicitly.

## 2.5. Description of the Free Energy Surface and Comparison to Experiment

---

harmonic oscillators is thus given by

$$E_v = \sum_{k=1}^{3M-6} \hbar\omega_k \left(n_k + \frac{1}{2}\right). \quad (2.187)$$

The ground level for which all quantum numbers are zero, *i.e.*  $n_1 = n_2 = \dots = n_{3M-6} = 0$ , gives the zero-point energy (ZPE) of the system:

$$E_v^{\text{ZPE}} = \frac{1}{2} \sum_{k=1}^{3M-6} \hbar\omega_k. \quad (2.188)$$

The levels for which one quantum number equals one ( $n_{\tilde{k}} = 1$ ) and all other quantum numbers are zero ( $n_k = 0$  if  $k \neq \tilde{k}$ ) are called the fundamental levels. The levels for which one quantum number is larger than one ( $n_{\tilde{k}} > 1$ ) and all other quantum numbers are zero ( $n_k = 0$  if  $k \neq \tilde{k}$ ) are called the overtone levels. The levels for which at least two quantum numbers have non-zero values are called the combination levels. The vibrational energy levels may be excited or de-excited by absorbing or emitting photons. Transition between levels takes only place if the energy of the photon matches the energy difference between the levels. For molecules, the absorption or emission spectrum arising from vibrational motion is mostly in the infrared (IR) region, *i.e.* the region of wave numbers from about  $200 \dots 4000 \text{ cm}^{-1}$ . Thus, absorption experiments that probe vibrational motion of molecules by passing light from a suitable source through a chamber containing the molecules to be studied are simply called IR spectroscopy experiments. The description of such an experiment that will be made use of in Chapter 3 of this thesis is provided in Subsection 2.5.2. From Equation (2.187) it is evidently clear that a transition from the ground level to a fundamental level will have the frequency  $\omega_{\tilde{k}}$  that is just the classical frequency of the  $\tilde{k}$ th normal mode. These so-called fundamental frequencies are commonly the most important ones in IR spectra because the ground level is usually the most populated one. The intensity of a spectral line is determined by the transition probability between the two vibrational levels. For estimation purposes, it is often justified to treat the interaction of the dipole moment of the molecule with the external electromagnetic field as a small perturbation to the system. Using Fermi's golden rule [300, 301], it is found that the IR intensity  $\mathcal{I}_{\text{IR}}$  is governed by the absolute square of the transition dipole moment  $(\vec{\mu})_{n,n'}$ , *i.e.*

$$\mathcal{I}^{\text{IR}} \sim |(\vec{\mu})_{n,\tilde{n}}|^2. \quad (2.189)$$

The two quantum numbers  $n$  and  $\tilde{n}$  thereby denote the two vibrational levels between which the transition takes place. Using the quantum numbers  $n_k$ ,  $k = 1, \dots, 3M - 6$ , associated with the  $3M - 6$  quantum harmonic oscillators in Equation (2.187), they are given by

$$\begin{aligned} n &= n_1 + n_2 + \dots + n_{3M-6} \\ \text{and } \tilde{n} &= \tilde{n}_1 + \tilde{n}_2 + \dots + \tilde{n}_{3M-6}. \end{aligned} \quad (2.190)$$

The transition dipole moment in Equation (2.189) is given by

$$(\vec{\mu})_{n,\tilde{n}} = \int (\Psi_{\nu}^*)_{n} \vec{\mu} (\Psi_{\nu})_{\tilde{n}} d\Omega_{\nu}, \quad (2.191)$$

where  $(\Psi_{\nu})_{n}$  and  $(\Psi_{\nu})_{\tilde{n}}$  denote the vibrational wavefunctions (see Equation (2.183)) corresponding to the vibrational levels denoted by the quantum numbers  $n$  and  $\tilde{n}$ , respectively. The integration is over the whole configuration space associated with the vibrations described by the normal coordinates  $Q_k$ ,  $k = 1, \dots, 3M - 6$ , for which  $d\Omega_{\nu}$  denotes an infinitesimal volume element. Assuming the external electromagnetic field as a small perturbation to the system, the dipole moment  $\vec{\mu}$  may then be expanded in a Taylor series in terms of the normal coordinates  $Q_k$ , *i.e.*

$$\vec{\mu} = \vec{\mu}_0 + \sum_{k=1}^{3M-6} \left( \frac{\partial \vec{\mu}}{\partial Q_k} \right)_0 Q_k + \dots \quad (2.192)$$

The “electrical linear approximation” now consists in truncating the expansion at first order. Together with the “mechanical harmonic approximation” in Equation (2.163), one commonly denotes the approach as “double harmonic approximation” [302, 303]. The transition dipole moment in Equation (2.191) is then given by

$$(\vec{\mu})_{n,\tilde{n}} = \vec{\mu}_0 \int (\Psi_{\nu}^*)_{n} (\Psi_{\nu})_{\tilde{n}} d\Omega_{\nu} + \sum_{k=1}^{3M-6} \left( \frac{\partial \vec{\mu}}{\partial Q_k} \right)_0 \int (\Psi_{\nu}^*)_{n} Q_k (\Psi_{\nu})_{\tilde{n}} d\Omega_{\nu}. \quad (2.193)$$

The first term considers the permanent dipole moment  $\vec{\mu}_0$  of the molecule and vanishes unless  $n = \tilde{n}$  due to the orthonormality of the vibrational wavefunctions. In other words, it does not effect the intensities of the vibrational spectrum. It can be shown that the second term does not vanish only if  $n = \tilde{n} \pm 1$  such that  $n_{\tilde{k}} = \tilde{n}_{\tilde{k}} \pm 1$  for a specific normal mode  $\tilde{k}$  and all other normal modes have  $n_k = \tilde{n}_k$  ( $k \neq \tilde{k}$ ) [297]. This obviously concerns fundamental frequencies  $\omega_{\tilde{k}}$  for which a change in the electric dipole moment of the molecule along the  $k$ th normal mode is caused, *i.e.* for which  $\left( \frac{\partial \vec{\mu}}{\partial Q_k} \right)_0 \neq 0$ . Only when going beyond the double harmonic approximation, overtone and combination transitions may be expressed as well. Using Equation (2.189) and Equation (2.193), the IR intensity  $\mathcal{I}_k^{\text{IR}}$  of the spectral line associated with the fundamental frequency  $\omega_k$  in double harmonic approximation is thus governed by the absolute square of the change in the electric dipole moment of the molecule along the  $k$ th normal mode, *i.e.*

$$\mathcal{I}_k^{\text{IR}} \sim \left| \left( \frac{\partial \vec{\mu}}{\partial Q_k} \right)_0 \right|^2. \quad (2.194)$$

In order to calculate the proportionality constant, one needs to rely on first-order time-dependent perturbation theory [304]. One yields

$$\mathcal{I}_k^{\text{IR}} = \frac{N_A \pi}{3c} \left| \left( \frac{\partial \vec{\mu}}{\partial Q_k} \right)_0 \right|^2, \quad (2.195)$$

where  $N_A$  denotes the Avogadro constant and  $c$  is the speed of light in vacuum.

## 2.5. Description of the Free Energy Surface and Comparison to Experiment

After having laid out the theoretical background, evaluating IR spectra in double harmonic approximation within the framework of DFT – here using the electronic structure code package FHI-aims – provides no further difficulty when ensuring a good estimate of the mass-weighted Hessian matrix  $\frac{\partial^2 V_{\text{BO}}}{\partial q_i \partial q_j}$  given in Equation (2.173). Within the framework of DFT, the Born-Oppenheimer potential  $V_{\text{BO}}$  in Equation (2.161) is expressed as (see Equations (2.32), (2.82), and (2.88))

$$V_{\text{BO}} = E_{\text{total}}^{\text{DFT}} = \frac{1}{2} \sum_{k_1 \neq k_2=1}^M \frac{Z_{k_1} Z_{k_2} e^2}{|\vec{R}_{k_1} - \vec{R}_{k_2}|} + \tilde{T}[\rho] + E_{\text{ext}}[\rho] + E_{\text{C}}[\rho] + E_{\text{xc}}[\rho], \quad (2.196)$$

where the first term denotes the internuclear repulsion,  $E_{\text{C}}[\rho]$  is the classical Coulomb energy of the electron density  $\rho$ ,  $\tilde{T}[\rho]$  is the Kohn-Sham kinetic energy,  $E_{\text{ext}}[\rho]$  is the external potential energy, and  $E_{\text{xc}}[\rho]$  is the xc energy functional. Calculation of mass-weighted atomic forces  $\vec{F}_k$ ,  $k = 1, \dots, M$ , *i.e.*

$$\vec{F}_k = -\frac{\partial E_{\text{total}}^{\text{DFT}}}{\partial \vec{q}_k}, \quad (2.197)$$

is thus straightforward but tedious, refer to Section 4.7 of Reference [264] for detailed description within FHI-aims. Making use of a small finite displacements approach [305], the mass-weighted Hessian matrix in Equation (2.173) is then estimated as

$$\frac{\partial^2 V_{\text{BO}}}{\partial q_i \partial q_j} \approx \frac{F_i(q_1, \dots, q_j + \Delta^{(k)}, \dots, q_{3M}) - F_i(q_1, \dots, q_j - \Delta^{(k)}, \dots, q_{3M})}{2\Delta^{(k)}}, \quad i, j = 1, \dots, 3M, \quad (2.198)$$

where the indices  $i, j$  have been used again to enumerate atomic coordinates. In other words, each atom  $k$  is displaced in three spatial directions by a small finite displacement  $\Delta^{(k)} = \sqrt{M_k} \delta$ , the forces are calculated at each displacement, and the mass-weighted Hessian matrix is estimated accordingly. Displacement values of  $\delta = 10^{-3} \text{ \AA} \dots 10^{-2} \text{ \AA}$  have been shown to give reliable results [306]. This approach requires  $6M + 1$  single-point energy calculations including force evaluations, meaning computational costs should be taken into consideration. Similarly, the change in the electric dipole moment  $\frac{\partial \vec{\mu}}{\partial q_i}$  needed for evaluating the IR intensity in Equation (2.195), is estimated as

$$\frac{\partial \vec{\mu}}{\partial q_i} \approx \frac{\vec{\mu}(q_1, \dots, q_i + \Delta^{(k)}, \dots, q_{3M}) - \vec{\mu}(q_1, \dots, q_i - \Delta^{(k)}, \dots, q_{3M})}{2\Delta^{(k)}}, \quad i = 1, \dots, 3M. \quad (2.199)$$

For any given set of (mass-weighted) atomic coordinates, the dipole moment  $\vec{\mu}$  of the molecule is calculated as [307]

$$\vec{\mu} = \sum_{k=1}^M Z_k \vec{R}_k + \int \rho(\vec{r}) \vec{r} d\vec{r}, \quad (2.200)$$

where the  $Z_k$  denote the net nuclear charges and the  $\vec{R}_k$  are the atomic positions. The second term is just the first moment of the electronic density. Although the dipole moment  $\vec{\mu}$  of the molecule depends on the choice of the origin of the coordinate system (except for neutral molecules), the same does not hold true for the dipole moment derivatives that enter the IR

intensity calculation in Equation (2.195).

Calculating free energy contributions using the concept of the Helmholtz free energy now provides no further difficulty. Following the decoupling *Ansatz* in Equation (2.181), the energy of the system is approximated as being separable into electronic, translational, vibrational, and rotational contributions. In other words, the Helmholtz free energy  $F$  per molecule is given by

$$F = E_e + F_{\text{int}}, \quad (2.201)$$

$$\text{with } F_{\text{int}} = F_{\text{trans}} + F_{\text{rot}} + F_{\text{vib}},$$

where the electronic energy  $E_e$  is obtained by solving the electronic Schrödinger equation (Equation (2.37)) for which a variety of theoretical models and levels of theory have been discussed in Subsections 2.3.1 through 2.3.8.  $F_{\text{int}}$  denotes the free energy contribution due to the internal degrees of freedom, consisting of translation, vibration, and rotation. The translational part of the free energy  $F_{\text{trans}}$  captures the impact of the pressure in a gas of the molecule [308]. However, as the goal of this work is to study peptide systems in the gas phase, *i.e.* in isolation, both the experiment described in Subsection 2.5.2 as well as the theoretical simulations are essentially done at zero pressure, hence justifying the neglect of the translational contribution to the Helmholtz free energy. Furthermore, throughout this work we are exclusively treating *relative* energies, *i.e.* comparing energy differences between different conformers (usually with respect to the global minimum) of the same system. Since the translational contributions only depend on the total molecular mass [126, 309], they will thus always cancel. Hence, the internal free energy  $F_{\text{int}}$  can be described in terms of its vibrational and rotational contributions only. As already discussed above, one thereby assumes neglect of any rotational-vibrational coupling. In other words, the rotation of the molecule is assumed to occur at fixed geometry, giving rise to the so-called rigid-rotor approximation. Within the framework of quantum statistical mechanics, the Helmholtz free energy  $F$  is formally defined by [310, 311]

$$F = -k_B T \ln Z, \quad (2.202)$$

where  $k_B$  denotes the Boltzmann constant,  $T$  is the temperature, and  $Z$  denotes the canonical partition function that is defined as

$$Z = \sum_i e^{-\epsilon_i / k_B T}, \quad (2.203)$$

where the sum is over all possible quantum energy states  $\epsilon_i$  of the system. For a rigid rotor, *i.e.* a rigid rotating polyatomic molecule, the corresponding canonical partition function  $Z_{\text{rot}}$  is given by [312]

$$Z_{\text{rot}} = \sqrt{\pi} \left( \frac{2k_B T}{\hbar^2} \right)^{3/2} \sqrt{I_1 I_2 I_3}, \quad (2.204)$$

where  $I_1$ ,  $I_2$ , and  $I_3$  denote the three different principal moments of inertia. The rotational



## 2.5. Description of the Free Energy Surface and Comparison to Experiment

Helmholtz free energy  $F_{\text{rot}}$  is thus given by

$$F_{\text{rot}} = -k_{\text{B}}T \ln \left[ \sqrt{\pi} \left( \frac{2k_{\text{B}}T}{\hbar^2} \right)^{3/2} \sqrt{I_1 I_2 I_3} \right]. \quad (2.205)$$

Treating the vibrational contributions in harmonic approximation is straightforward as the possible vibrational quantum energy states of the system of a set of  $3M - 6$  coupled quantum harmonic oscillators are known from Equation (2.187). The corresponding canonical partition function  $Z_{\text{vib}}$  is thus given by

$$Z_{\text{vib}} = \prod_{k=1}^{3M-6} \sum_{n_k=0}^{\infty} e^{-\frac{\hbar\omega_k}{k_{\text{B}}T}(n_k+\frac{1}{2})} = \prod_{k=1}^{3M-6} \frac{e^{-\frac{\hbar\omega_k}{2k_{\text{B}}T}}}{1 - e^{-\frac{\hbar\omega_k}{k_{\text{B}}T}}}, \quad (2.206)$$

where the  $\omega_k$  again denote the classical normal frequencies of the system. Inserting Equation (2.206) into Equation (2.202) yields

$$F_{\text{vib}} = \sum_{k=1}^{3M-6} \left[ \frac{\hbar\omega_k}{2} + k_{\text{B}}T \ln \left( 1 - e^{-\frac{\hbar\omega_k}{k_{\text{B}}T}} \right) \right]. \quad (2.207)$$

For  $T = 0$ , the system exists in its ground state and the internal Helmholtz free energy  $F_{\text{int}}$  gives the zero-point energy (ZPE) of the system, *i.e.*  $F_{\text{int}} = \frac{1}{2} \sum_{k=1}^{3M-6} \hbar\omega_k$ , as already derived in Equation (2.188).

### 2.5.2 Experimental Setup

The comparison of calculated vibrational spectra derived from molecular simulations as described in the previous subsection with experimentally observed IR spectra helps to characterize structural motifs of peptides and allows for structure elucidation. On one hand, theoretical predictions help to interpret experimentally obtained spectra. On the other hand, a rigorous experiment-theory comparison allows for the assessment of the accuracy and predictive power of simulation approaches. A detailed description of the experimental setup that will be referred to extensively in Chapter 3 of this thesis is provided in Reference [313] and will be briefly summarized in the following.

The machine used for performing spectroscopic studies of peptide systems in the gas phase is a cold-ion spectroscopy instrument for which a schematic illustration is shown in Figure 2.10. It combines a nano-electrospray ion source with a cryogenic octopole ion trap ( $T = 4\text{K}$ ) and allows for performing IR-UV double resonance spectroscopy [315] in order to obtain conformer-selective vibrational spectra. In brief, positively charged gas-phase peptides are produced in a continuous fashion by nano-electrospray ionization from a 0.1 mM solution in 50:50 methanol-water. After entering the instrument through a metal-coated borosilicate capillary, the protonated peptides are focused by an ion funnel. The peptides are pre-trapped in a hexapole in order to generate ion packets and to match the duty cycle of the experiment. A

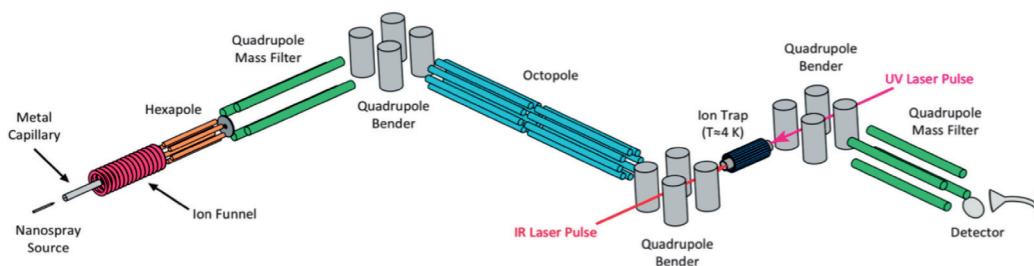


Figure 2.10 – Schematic illustration of the cold-ion spectroscopy instrument. Reproduced with permission from Reference [314].

quadrupole mass filter selects the mass-to-charge ratio of the peptides of interest, after which they are deflected  $90^\circ$  using a quadrupole bender. The charged molecules are guided through an octopole and deflected  $90^\circ$  a second time before passing through a set of decelerating lenses. Finally, they are injected into the cold octopole ion trap ( $T = 4\text{ K}$ ). Here, they are cooled down to approximately  $10\text{ K}$  by collisions with cold helium gas of pressure  $6 \cdot 10^{-6} \dots 10^{-5}\text{ mbar}$  that is pulsed in before their arrival. Infrared (IR) and ultraviolet (UV) beams are focused inside the trap and used to spectroscopically interrogate the cold molecules. Following UV absorption of the parent ions, the produced charged fragments are extracted from the trap, deflected by a third electrostatic bender, and passed through a quadrupole mass filter which selects a particular mass-to-charge ratio before they are detected by a channeltron electron multiplier. The electronic signature of the protonated peptides is recorded monitoring the number of fragments for a particular photofragmentation channel as a function of the UV wavenumber. Each conformer has a characteristic UV signature, meaning the recorded spectrum is a superimposition of lines coming from all conformations of the parent peptide that may be present in the trap. Fixing the wavenumber of the UV laser and scanning the wavenumber of an infrared laser pulse that arrives  $200\text{ ns}$  earlier allows for acquiring a vibrational spectrum of whatever conformer is resonant with the UV laser. When the IR pulse is in resonance with a vibrational transition of the ion, part of the population is removed from the ground state, thus leading to a decrease in UV induced fragmentation. Hence, as the IR wavenumber is scanned, one obtains a conformer-specific vibrational spectrum. Performing the same procedure on each line of the electronic spectrum allows for assigning each UV spectral feature to a particular conformer.

### **3 Conformational Structure Search and Kinetically Trapped Liquid-State Conformers of a Sodiated Model Peptide Observed in the Gas Phase**

Results and findings described in this chapter have been published in Reference [316] [Schneider *et al.* *J. Phys. Chem. A*, **121**, 6838-6844 (2017)] and are being reproduced here.

### 3.1 Motivation: Prerequisites of Helix Formation in the Gas Phase

Helices are common secondary structural motifs in peptides and proteins. As explained in Section 2.2, there exist several helix types that can be characterized based on the intramolecular hydrogen bond patterns of the backbone alone, with  $\alpha$ -helices and  $3_{10}$ -helices being the most common types [54, 317]. In solution, helix propensity is determined both by intramolecular interactions and protein-solvent interaction. In this work however, the focus lies on peptide systems in the gas phase as the main goal is to study intramolecular interactions of peptides in isolation. As laid out in Section 2.2, this is because gas-phase systems offer the opportunity to study the “undamped” intramolecular interactions that shape peptides, thereby shedding light on intrinsic structural motif propensities and bonding interactions. Gas-phase helices have been investigated using ion mobility spectrometry [318–320] and vibrational spectroscopy [321–328]. The combination of these experimental techniques with molecular simulations based on DFT allows for structure elucidation, as it helps to interpret experimentally obtained spectra. Moreover, a rigorous experiment-theory comparison allows for the assessment of the accuracy and predictive power of simulation approaches [329].

Pioneering ion-mobility experiments in the group of Jarrold [318, 319] examined the role of N- and C-terminal residues on gas-phase helix formation for the sequences  $\text{Ala}_n\text{H}^+$ ,  $\text{AcLysAla}_n\text{H}^+$ , and  $\text{AcAla}_n\text{LysH}^+$ . They concluded that  $\text{Ala}_n\text{H}^+$  and  $\text{AcLysAla}_n\text{H}^+$  adopt globular conformations in the gas phase independent of the length of the amino-acid chain while  $\text{AcAla}_n\text{LysH}^+$  is helical for  $n > 8$  [330]. The identities of these structures were confirmed by theoretical and experimental vibrational spectroscopy in the work of Rossi *et al.* [326] and Schubert *et al.* [328]. Similar studies focused on peptides of the form  $\text{AcPheAla}_n\text{LysH}^+$  with  $n = 1-5, 10$ , where phenylalanine (Phe) provides a UV chromophore, which allows for conformer-specific IR-UV double resonance spectroscopy [322–325], as described in Subsection 2.5.2. In these experiments, the number of residues necessary to form a helix was found to be six [324, 330], but much of the hydrogen bonding pattern responsible for the formation of this motif is already present even with only three residues [325, 331]. In conjunction with computational vibrational spectroscopy based on DFT [324–326, 332], such spectra allowed for determining detailed molecular structures and critically examining evidence for helix formation of peptides in isolation.

The helix-stabilizing factors in polyalanine peptides are illustrated in Figure 3.1 for the specific case of  $\text{AcPheAla}_5\text{LysH}^+$ . Work by the groups of Jarrold [318, 319], Rizzo [322–325], and Blum [328] showed that intramolecular hydrogen bonds play an important role and that the design concept can even be transferred to non-natural peptides [327]. Hoffmann *et al.* [333] could show that deleting a single hydrogen bond had little impact on the overall helix stability. In addition to their energetic stability, hydrogen bonds are aligned in helices, and the resulting

### 3.1. Motivation: Prerequisites of Helix Formation in the Gas Phase

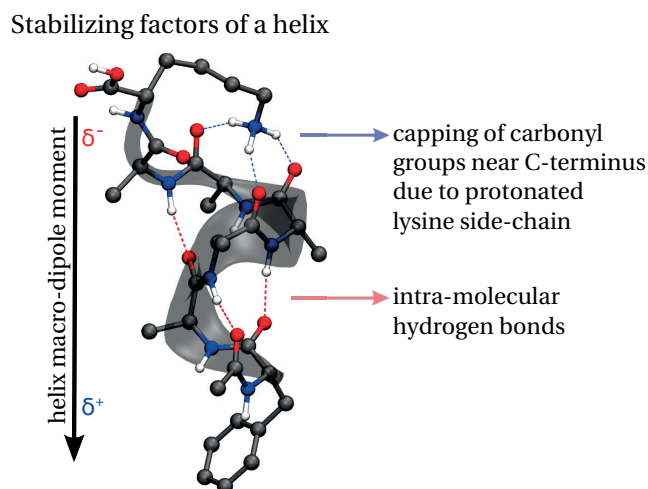


Figure 3.1 – Illustration of helix-stabilizing factors for peptides in the gas phase for the specific case of AcPheAla<sub>5</sub>LysH<sup>+</sup>.

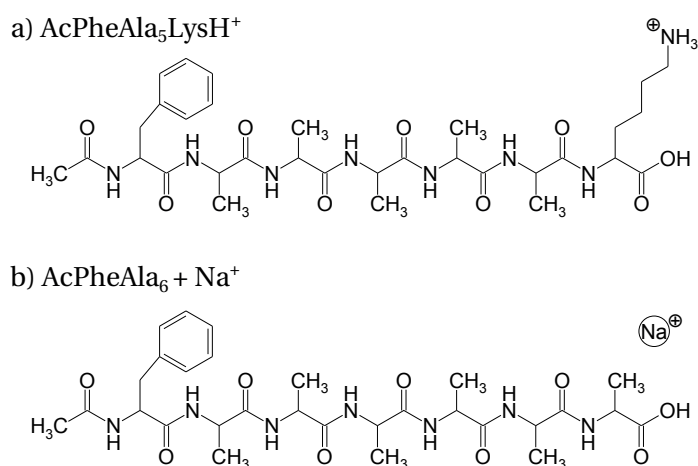


Figure 3.2 – Structural formulas of (a) AcPheAla<sub>5</sub>LysH<sup>+</sup> and (b) AcPheAla<sub>6</sub> + Na<sup>+</sup>.

macro-dipole favorably interacts with the positive charge of the protonated lysine (Lys) side-chain at the C-terminus. Moreover, the capping of the “dangling” carbonyl groups near the C-terminus by the Lys side-chain provides additional stability.

In order to obtain a more complete picture, the importance of the charge fixed at the C-terminus is investigated. To that end, the focus lies on the well-studied system [324, 332] of AcPheAla<sub>5</sub>LysH<sup>+</sup> which is compared to AcPheAla<sub>6</sub> + Na<sup>+</sup>. The structural formulas of both systems are provided in Figure 3.2. In the latter, Lys is formally replaced by alanine (Ala) and a sodium cation (Na<sup>+</sup>) in order to introduce a freely movable positive charge. The resulting rich possibilities for electrostatic interaction can locally disrupt hydrogen-bonding networks and induce unconventional backbone conformations [206, 334–336]. Consequently, the cation-

## Chapter 3. Kinetically Trapped Conformers of a Sodiated Peptide in the Gas Phase

---

binding site, and hence the conformation as a whole, is not *a priori* obvious. Ion mobility studies on metallated peptides, *e.g.* sodiated species of  $\text{Ala}_n + \text{M}^+$  [337], suggest that the cation plays the same role as the charged Lys side-chain in  $\text{AcAla}_n\text{LysH}^+$  for peptides with  $n > 12$ . For shorter peptides, calculated collisional cross sections (CCS) for globular and helical structures are both in agreement with the experimental CCS, preventing a definitive structural assignment. In this work, IR-UV double resonance spectroscopy and theory are coupled in order to unravel the structure of the system of  $\text{AcPheAla}_6 + \text{Na}^+$  with the aim of understanding whether a freely movable cation is sufficient to stabilize helix formation or if the C-terminal localization is a prerequisite for that.

### 3.2 Experimental Setup

The experimental setup has already been described in Subsection 2.5.2. In brief, a nano-electrospray ion source is combined with a cooled ion trap for spectroscopic studies of gas-phase ions. Conformer-selective IR spectra are recorded by applying IR-UV double resonance. A measurement is performed by fixing the wavenumber of the UV laser to a line in the electronic spectrum and scanning the wavenumber of an infrared laser. When the IR pulse is in resonance with a vibrational transition of the ion, part of the population is removed from the ground state, leading to a decrease in UV-induced fragmentation. Scanning the IR wavenumber, one obtains a conformer-specific vibrational spectrum. Performing the same experiment on each line of the electronic spectrum allows for assignment of each UV spectral feature to a particular conformer.

### 3.3 Computational Methods

The applied conformational search algorithm is similar to the one used by Rossi *et al.* [332]. First, a global conformational search is performed on the force field (FF) level (refer to Subsection 2.3.1 for details) using the two empirical fixed point charge models of CHARMM22 and OPLS-AA, separately. To that end, the basin-hopping approach described in Section 2.4 is applied using the `scan` program of the TINKER molecular modeling package [260]. To be detailed, *all* torsional modes are taken into consideration and default search parameters are used, *i.e.* an energy threshold for local minima of 100 kcal/mol and a convergence criterion for local geometry optimizations of 0.0001 kcal/mol·Å. For the system of  $\text{AcPheAla}_5\text{LysH}^+$ , 603 280 conformers are found using CHARMM22, and 643 938 conformers are found using OPLS-AA. For the system of  $\text{AcPheAla}_6 + \text{Na}^+$ , 626 829 conformers are found using CHARMM22, and 635 120 conformers are found using OPLS-AA. All subsequent DFT calculations are done using the electronic structure code package FHI-aims for which computational details have been described in Subsection 2.3.9. Single-point energy calculations on the PBE+vdW<sup>TS</sup> level of DFA (refer to Subsections 2.3.5 and 2.3.6 for details) using `tier 1` basis sets and `light` settings are performed for *all* these FF conformers. For the two FFs individually, the 500 conformers with the lowest FF energy and the 500 conformers with the lowest DFT energy, *i.e.* a

grand total of 2000 conformers, are selected. The 2000 selected conformers are then geometry optimized at the PBE+vdW<sup>TS</sup> level using `tier 1` basis sets and `light` settings. Relaxation is accomplished using a trust radius method version of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm [338]. After convergence, a clustering scheme is applied in order to rule out duplicates. To be precise, root-mean-square deviations (RMSD) of atomic positions between any two conformers are calculated using OpenBabel [339]. Hierarchical clustering is then achieved by applying the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [340] method implemented in Python's SciPy [341] library. Following that, further relaxation is accomplished at the PBE+vdW<sup>TS</sup> level using `tier 2` basis sets and `tight` settings. After clustering, this results in 324 conformers for AcPheAla<sub>5</sub>LysH<sup>+</sup> and 159 conformers for AcPheAla<sub>6</sub> + Na<sup>+</sup> in the low-energy region, *i.e.* within 6 kcal/mol from the global minimum. These conformers are then again locally refined on the PBE0+MBD level using `tier 1` basis sets and `light` settings. After clustering, further geometry relaxation on the PBE0+MBD level using `tier 2` basis sets and `tight` settings results in 52 conformers for AcPheAla<sub>5</sub>LysH<sup>+</sup> and 23 conformers for AcPheAla<sub>6</sub> + Na<sup>+</sup> in the low-energy region, *i.e.* within 3 kcal/mol from the global minimum.

## 3.4 Results and Discussion

### 3.4.1 AcPheAla<sub>5</sub>LysH<sup>+</sup>

For the present comparative study, a firm assignment of measured conformer-selective IR spectra to their calculated counterparts is of paramount importance. To that end, the peptide AcPheAla<sub>5</sub>LysH<sup>+</sup> is being re-assessed first, thereby demonstrating that the applied conformational search technique completely grasps the conformational space energetically close to the global minimum, and that the applied level of theory is capable of reproducing the energetics as well as the vibrational properties of the conformers. For this, the results are compared to previous work on AcPheAla<sub>5</sub>LysH<sup>+</sup> by Stearns *et al.* [324], where the 45 lowest-energy structures were selected out of a set of 1,000 force-field minima and subsequently optimized using DFA with a hybrid xc functional. Even though four structures were successfully assigned to the experimental spectra, the question whether the search was complete and whether these conformers are located in the global minimum region remained open. This did, in part, motivate an exhaustive conformational search by Rossi *et al.* [332], in which 7 conformers were found within 1 kcal/mol of the global minimum on the PES. The authors were able to assign the experimentally observed structures to the global minima populated at low temperature by using the hybrid xc functional PBE0 augmented by the MBD correction and including zero-point energy corrections. The latter were computed with the GGA functional PBE and the pair-wise Tkatchenko-Scheffler van der Waals correction (vdW<sup>TS</sup>), which proved however unsatisfying for the prediction of vibrational spectra. It was suggested that using a hybrid xc functional was necessary, which was a natural assumption since this level of theory was necessary for a correct conformational energy prediction in the first place. Furthermore, it was assumed that an anharmonic treatment was needed to yield improved spectra.

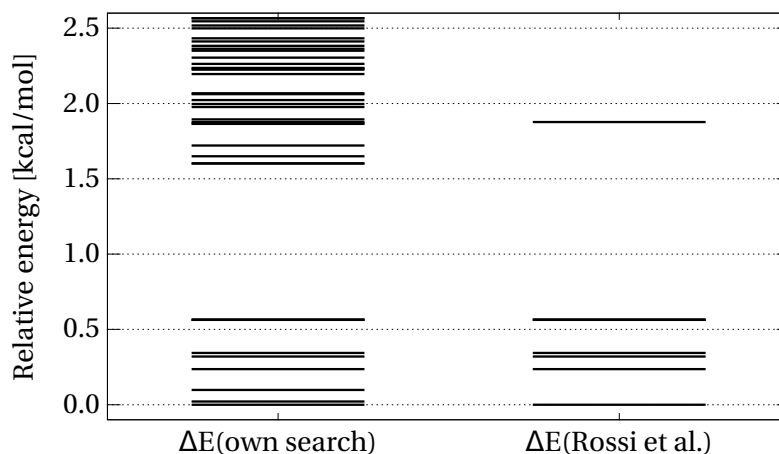


Figure 3.3 – Comparison of energy hierarchies of conformers of AcPheAla<sub>5</sub>LysH<sup>+</sup> at the PBE0+MBD level (tier 2 basis sets and tight settings) between the conformational search applied here and the search performed by Rossi *et al.* [332]. Two additional conformers are found in the low-energy region, *i.e.* within 1 kcal/mol from the global minimum. Conformers with the same energy in both hierarchies correspond to virtually identical structures.

The conformational search strategy has already been laid out in detail in Section 3.3, including numbers illustrating the exhaustiveness of the search. The fact that two additional conformers are found within 1 kcal/mol from the lowest-energy conformer gives confidence in the conformational search. Figure 3.3 compares the two corresponding hierarchies of the relative DFT energy  $\Delta E$  on the PES, *i.e.* on the PBE0+MBD level using tier 2 basis sets and tight settings. Conformers with the same energy in both hierarchies correspond to virtually identical structures. In total, nine conformers were found within 1 kcal/mol from the global minimum.

Since the experimental measurement takes place on cold ions in the gas phase, the PES merely allows for a rough estimate about the structures populated at low temperatures. To confidently assign the experimentally observed structures one needs to rely on the Helmholtz free energy  $F$  at 10 K because this is approximately the temperature of the observed ions, as explained in Subsection 2.5.2. Free energy contributions are accounted for from internal degrees of freedom, consisting of vibrations and rotations, in addition to the DFT energy  $E$  on the PES. A detailed formulaic description is provided in Subsection 2.5.1, see Equations (2.201) through (2.207). For AcPheAla<sub>5</sub>LysH<sup>+</sup>, Figure 3.4 shows energy hierarchies of the PBE0+MBD energy  $\Delta E$  as well as the Helmholtz free energy  $\Delta F$  at 10 K and at 300 K, always relative to conformer **A** (see Figure 3.5(b)). At this stage, harmonic vibrational free energy contributions have been calculated at the PBE+vdW<sup>TS</sup> level. While the  $\Delta F(10\text{ K})$  surface should best resemble experimental conditions of gas-phase measurements at 10 K, the free energy hierarchy at 300 K represents an estimate of the conformers populated at the early stage of the experimental process, where the molecules are electrosprayed into the instrument at room temperature. Their low free energy at 10 K and the relatively large gap to alternative structures at 300 K



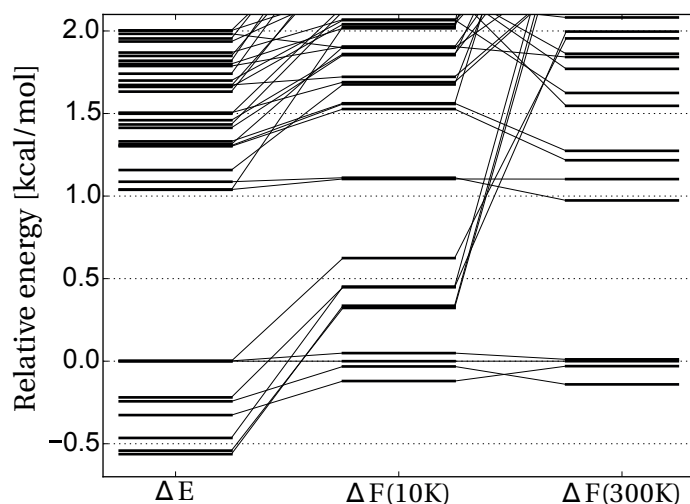


Figure 3.4 – Energy hierarchies of conformers of AcPheAla<sub>5</sub>LysH<sup>+</sup> at the PBE0+MBD energy  $\Delta E$  as well as the Helmholtz free energy  $\Delta F$  at 10K and 300K with harmonic vibrational free energy contributions calculated at the PBE+vdW<sup>TS</sup> level.

indicate why the species observed in experiment should be among the four conformers within 0.25kcal/mol from the global minimum. Of course, one needs to be aware of the limitation of not taking into account anharmonicity and the possibility of solvation-memory effects (*i.e.* kinetic trapping).

High computational costs prohibited the systematic use of hybrid xc functionals for the calculation of harmonic vibrations in the previous study by Rossi *et al.* [332]. To complete the picture, the harmonic vibrational free energy calculations at the PBE0+MBD level are repeated, confirming the already obtained result. Figure 3.5(a) shows the energy hierarchies for  $\Delta E$ ,  $\Delta F(10K)$ , and  $\Delta F(300K)$  for the four lowest-energy conformers illustrated in Figure 3.5(b). Conformers **A** and **B** are virtually identical near the C-terminus, but differ near the N-terminus by a tilted Phe side chain. The difference between conformers **C** and **D** is similar. All four conformers show helical structure motifs: conformer **C** possesses one  $3_{10}$ - and two  $\alpha$ -helical turns, conformer **D** features one  $3_{10}$ - and one  $\alpha$ -helical turn, and conformers **A** and **B** each possess two  $3_{10}$ - and one  $\alpha$ -helical turn.

For this work, the original IR-UV double resonance experiment by Stearns *et al.* [324] has been repeated to allow conformer-selective IR spectra to be compared to their theoretical counterparts calculated at the PBE0+MBD level. The affiliated UV spectrum is provided in Figure 3.6 where peaks have been assigned to their identified conformers shown in Figure 3.5. The conformer-selective IR spectra that have been calculated in double harmonic approximation as explained in Subsection 2.5.1 are shown in Figure 3.5(c). Conformers **A** and **B** could be attributed to their corresponding observed IR spectra. While the agreement is very good, the match between experimental and theoretical IR spectra is not perfect. Reasons for this discrepancy have been touched in Subsection 2.5.1 concerning the limitations of

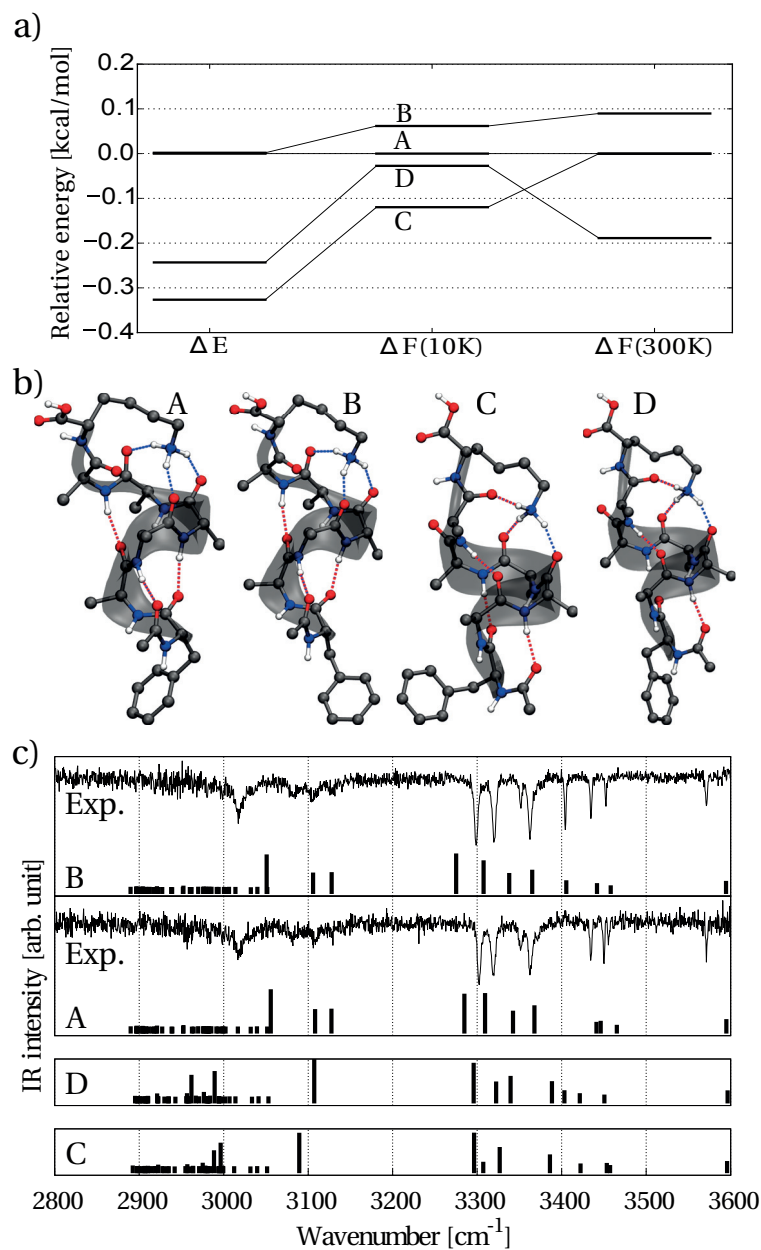


Figure 3.5 – (a) Relative DFT energies  $\Delta E$  as well as relative Helmholtz free energies  $\Delta F$  at 10 K and 300 K for the lowest-energy conformers of AcPheAla<sub>5</sub>LysH<sup>+</sup> at the PBE0+MBD level. (b) The four lowest-energy conformers on the  $\Delta F(10\text{ K})$  scale. Hydrogen bonds are indicated with dashed lines. The labeling of the conformers follows Stearns *et al.* [324]. (c) Two measured conformer-selective IR spectra (*traces*) are compared to double harmonic vibrational calculations (*sticks*). Calculated spectra were uniformly scaled by a factor of 0.948.

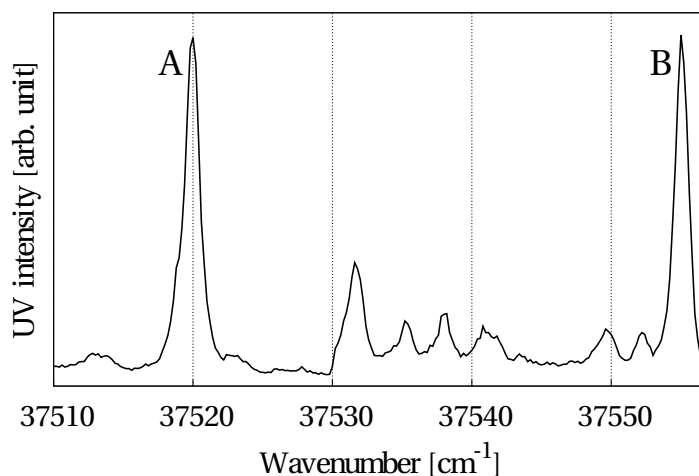


Figure 3.6 – Measured UV spectrum for the system of AcPheAla<sub>5</sub>LysH<sup>+</sup>. Peaks have been assigned to their identified conformers shown in Figure 3.5.

the double harmonic approximation approach. In summary, the discrepancy is commonly attributed to two factors: (i) The effect of a possible incomplete characterization of electron exchange and correlation, despite the use of the hybrid xc functional PBE0, and (ii) the treatment of anharmonic vibrations and nuclear quantum effects [342]. Both of these effects are corrected for solely by applying a scale factor to the vibrational frequencies. The assumption of a *uniform* overestimation of the harmonic vibrational modes with respect to experiment is debatable as they depend on the theoretical method, the used basis set, and the system itself [343, 344]. In this work, the focus lies on the frequency region of 3200 cm<sup>-1</sup> to 3500 cm<sup>-1</sup> which is sensitive to N–H···O hydrogen bonding, where a uniform scaling factor of 0.948 yields very good agreement.

The exhaustive conformational search presented here for AcPheAla<sub>5</sub>LysH<sup>+</sup>, and the rigorous treatment of harmonic vibrations at the hybrid xc level allowed for (i) reproducing the known energy hierarchy and finding additional conformers in the low-energy region and (ii) calculating well-fitting harmonic IR spectra for the conformers in the low-energy region. In this way, the conformers predicted by Stearns *et al.* [324] and Rossi *et al.* [332] are confirmed, and any other competing conformers can be ruled out. This also shows that calculating computationally costly anharmonic IR spectra is not required in this case. Now that the accuracy of the simulation approach has been confirmed, AcPheAla<sub>6</sub> + Na<sup>+</sup> is tackled, a more challenging system because of the additional conformational degrees of freedom due to the “unfixed” cation.

### 3.4.2 AcPheAla<sub>6</sub> + Na<sup>+</sup>

Figure 3.7 shows the energy hierarchies of the relative PBE0+MBD energies  $\Delta E$  as well as the relative Helmholtz free energies  $\Delta F$  at 10K and 300K with harmonic vibrational free

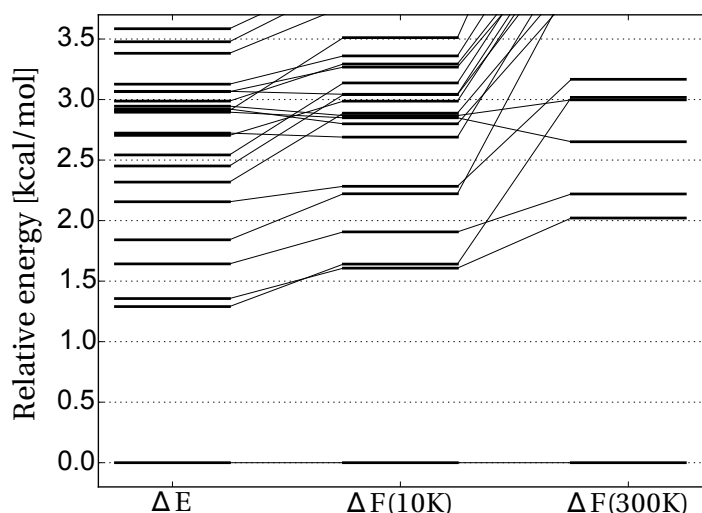


Figure 3.7 – Energy hierarchies of conformers of AcPheAla<sub>6</sub> + Na<sup>+</sup> at the PBE0+MBD energy  $\Delta E$  as well as the Helmholtz free energy  $\Delta F$  at 10K and 300K with harmonic vibrational free energy contributions calculated at the PBE+vdW<sup>TS</sup> level.

energy contributions at the PBE+vdW<sup>TS</sup> level that were obtained for AcPheAla<sub>6</sub> + Na<sup>+</sup>. The four presumably dominant conformers are presented in Figure 3.8(b). Of the four conformer-selective IR spectra that were recorded, two of them correspond to conformers with particularly high intensity in the UV spectrum, see Figure 3.9. The measured IR spectra of these two conformers, **IIa** and **IIb**, show very good agreement with the IR spectra calculated at the PBE0+MBD level, where again a scale factor of 0.948 has been applied. Both conformers are nearly identical, differing only in the tilt of the Phe side chain near the N-terminus. They are globular with the peptide being “wrapped around” the Na<sup>+</sup> cation with four partially negatively charged C=O groups pointing towards the positively charged cation, restricting them from forming the hydrogen bonds necessary for helix formation. Indeed, no similarities are observed comparing these structures to the helical motifs of AcPheAla<sub>5</sub>LysH<sup>+</sup>. The C-terminal fixation of the charge by the Lys side-chain seems to be a prerequisite to effectively cap the helix. The “freely movable” charge prevents helix formation in this system and instead induces a globular motif. All conformers found in the low-energy region (*i.e.* within 3 kcal/mol from the global minimum) show a globular conformation.

An obvious observation is the outstanding global minimum (conformer **I** in Figure 3.8(b)) that is separated by a 1.6 kcal/mol gap from the next minimum on the  $\Delta F(10K)$  scale. The clear assignment of conformers **IIa** and **IIb** to the two most intense bands in the measured spectra suggests that both conformers may be kinetically trapped. Moreover, the most stable structure **I** does not seem to be observed in the experiment – none of the conformer-selective spectra fit the calculated vibrational signatures (see Figure 3.8(c)). The structure representing the global minimum is globular and features a cation- $\pi$  interaction between the Na<sup>+</sup> and the Phe side chain. If that conformer were present in experiment, one would expect broad features in the

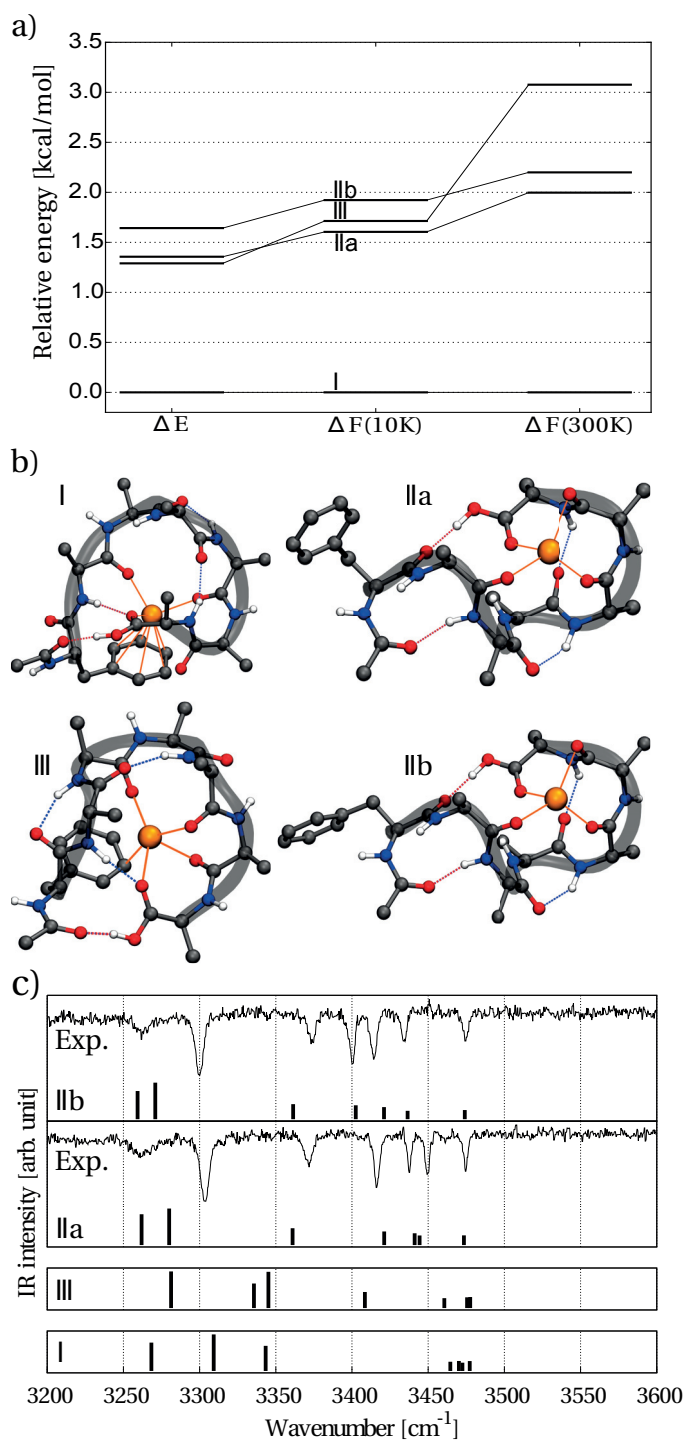


Figure 3.8 – (a) Relative DFT energies  $\Delta E$  as well as relative Helmholtz free energies  $\Delta F$  at 10 K and 300 K for the lowest-energy conformers of AcPheAla<sub>6</sub> + Na<sup>+</sup> at the PBE0+MBD level. (b) The four lowest-energy conformers on the  $\Delta F(10K)$  scale. Hydrogen bonds are indicated with dashed lines. The labeling of the conformers follows Stearns *et al.* [324]. (c) Two measured conformer-selective IR spectra (*traces*) are compared to double harmonic vibrational calculations (*sticks*). Calculated spectra were uniformly scaled by a factor of 0.948.

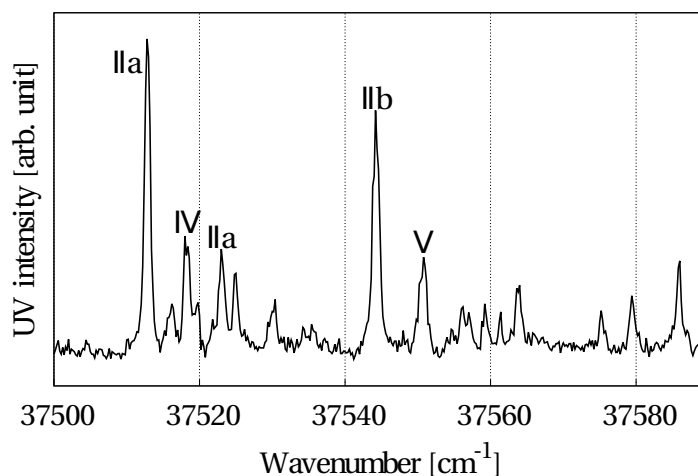


Figure 3.9 – Measured UV spectrum for the system of AcPheAla<sub>6</sub> + Na<sup>+</sup>. Peaks have been assigned to their identified conformers shown in Figure 3.8.

UV spectrum due to charge-transfer between Na<sup>+</sup> and the aromatic ring. However, no such features have been observed. The reason behind the kinetic trapping of conformers **IIa** and **IIb** has to be sought in the experimental procedure in which the molecules are electrosprayed into the apparatus from a solution at room temperature while the actual measurements are taken on isolated molecules at 10K. It is obvious from comparing the  $\Delta F(10\text{K})$  and  $\Delta F(300\text{K})$  hierarchies (see Figure 3.8(a)) that the temperature difference does not contribute to a possible kinetic trapping effect. In fact, the energy gap between the global and the next minimum even increases from 1.6kcal/mol at 10K to 2.0kcal/mol at 300K. Therefore, kinetic trapping must be caused by solvation effects. In order to estimate the magnitude of such an effect, re-relaxation was applied for the four lowest-energy conformers presented in Figure 3.8 On the PES at the PBE0+MBD level including implicit solvation effects by solving the Modified Poisson-Boltzmann (MPB) equation [345, 346] implemented [347] in FHI-aims. Default parameters have been chosen while explicitly setting `ions_conc 0` (no ions in the electrolyte). Full relaxation has been achieved for all conformers. Corresponding minima are still fairly similar as the root-mean-square deviation of atomic positions is smaller than 0.5Å in all cases. While in the gas phase conformer **I** is 1.6kcal/mol lower in DFT energy than the next minima (conformers **IIa** and **IIb**), the situation is reversed when including implicit aqueous solution; conformer **I** is now 0.9kcal/mol higher in energy. The situation is illustrated in Figure 3.10. This suggests that they carry a structural bias from aqueous solution, *i.e.* the barriers are sufficiently high to kinetically trap them during the electrospray process.

A similar scenario can be seen for conformer **III**, which is of comparable energy as conformers **IIa** and **IIb** on the  $\Delta F(10\text{K})$  scale, but the calculated IR spectrum, presented in Figure 3.8(c), does not match any experimentally observed one. Consulting the  $\Delta F(300\text{K})$  scale (see Figure 3.8(a)) shows that conformer **III** is 0.9kcal/mol higher in energy than conformer **IIb** at room temperature. When re-relaxing the structures to the nearest minimum on the PES at the

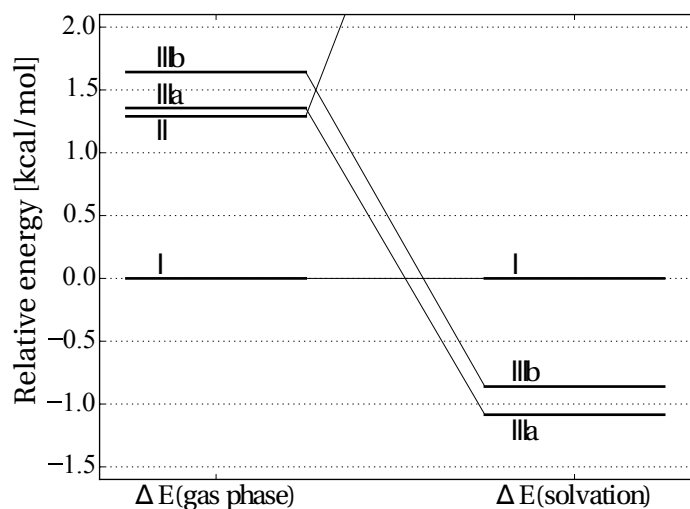


Figure 3.10 – Comparison of energy hierarchies on the PES at the PBE0+MBD level between gas-phase calculations and calculations including implicit solvation effects by solving the Modified Poisson-Boltzmann equation (MPB) implemented in FHI-aims. Full relaxation has been achieved for all conformers. Conformers have been labeled as in Figure 3.8. On the  $\Delta E(\text{solvation})$  scale conformer **III** lies 5.0 kcal/mol higher in energy than conformer **I**.

PBE0+MBD level including implicit aqueous solvation effects as described above, conformer **III** becomes further energetically penalized – it is then more than 5.0 kcal/mol higher in energy compared to the other conformers, as illustrated in Figure 3.10.

There remain two conformers, **IV** and **V**, for which the UV spectral signatures have lower intensity (see Figure 3.9), suggesting that they have smaller populations. The corresponding IR spectra, shown in Figure 3.11(a), could not be assigned to their calculated counterparts for any structure within 6 kcal/mol from the global minimum on the  $\Delta F(10\text{K})$  scale. Similarly, as for **IIa** and **IIb**, it is assumed that these conformers are kinetically trapped, which also renders their assignment difficult as these conformers might be higher in energy, and thus no energy criterion can be applied for finding them. Instead an approach [348] is followed where one makes use of information from the experiment in order to select from the overall pool of structures for calculation of spectra. Candidates were picked if they feature a free carboxylic acid OH stretch, since the experimental IR spectra show a peak at  $3578\text{ cm}^{-1}$  (see Figure 3.11(a)). Due to the absence of broad features in the UV spectrum, only structures were considered where the  $\text{Na}^+$  cation was not in close proximity to the phenyl ring. In total, vibrational spectra for 126 conformers have been calculated. In addition to that, local refinement at the PBE0+MBD level for all 52 found minima structures within 3 kcal/mol from the global minimum for the system of AcPheAla<sub>5</sub>LysH<sup>+</sup> has been laid out after formally replacing Lys with Ala + Na<sup>+</sup>, with the sodium cation being placed at the position of the amino group nitrogen. Vibrational spectra for the resulting 28 conformers (after clustering) have been calculated as well. As explained above, computationally-costly hybrid xc functionals are required in

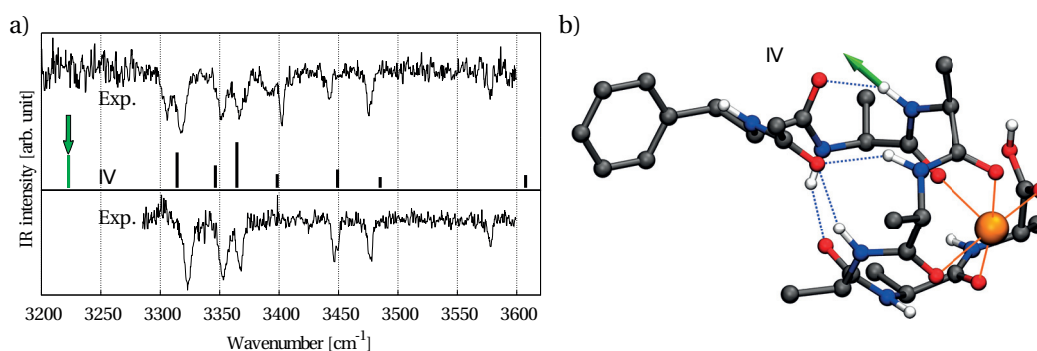


Figure 3.11 – (a) For the system of AcPheAla<sub>6</sub> + Na<sup>+</sup>, the two measured conformer-selective IR spectra (*traces*) with lowest intensity are compared to vibrational calculations (*sticks*) in double harmonic approximation at the PBE0+MBD level for structure **IV**. Calculated spectra have been scaled by applying a uniform scaling factor of 0.948. (b) Structural form of conformer **IV**. Hydrogen bonds are indicated with dashed lines. The highlighted vibrational mode in Figure (a) is indicated with a green arrow in Figure (b).

order to gain enough accuracy. Only conformer **IV** (see Figure 3.11(b)), lying 13.6 kcal/mol higher in energy than the global minimum on the  $\Delta F(10\text{K})$  scale, could be assigned to one of the less populated conformers. However, one peak in the simulated vibrational spectrum is blue shifted by  $80\text{ cm}^{-1}$  with respect to the nearest experimental peak, and the corresponding vibrational mode is indicated in Figure 3.11(b) with a green arrow. Conformer **IV** is a candidate for the kinetically trapped structure only because of the (partially) matching IR spectra. Taking into account the large computational effort taken, a more appropriate and computationally affordable technique for finding kinetically trapped conformers would be certainly desirable.

### 3.5 Conclusion

The data indicates that the fixed location of the charge at the C-terminus is imperative for helix formation in peptides of this length in isolation, as this stabilizes the structure through a cation-helix dipole interaction. In the case of the freely-movable sodium cation, the cation-backbone and cation- $\pi$  interactions seem to be stronger, leading to local distortions of peptide structure, preventing helix stabilization. It is interesting to note the high barriers that seem to be involved in interconverting one structure to another. Even though the cation- $\pi$  interaction is energetically favored for AcPheAla<sub>6</sub> + Na<sup>+</sup> in the gas phase, the system remains kinetically trapped in a structural state that is characterized by cation-backbone interactions and that is energetically preferred in polar solvent.



## **4 Energetics and Benchmark of Across-the-scale Energy Methods of Acetyl-Histidine Protomers with and without Zn<sup>2+</sup>**

Results and findings described in this chapter have been collected in a manuscript and are about to be submitted [349].

### 4.1 Motivation and Overview

Metal cations often play a crucial role in shaping the three-dimensional structure of proteins and peptides. Examples of significant conformational changes of peptides in their presence that may alter important properties were presented in Chapter 1, see *e.g.* Figure 1.1. It is obviously much desirable to have a very good fundamental and detailed theoretical understanding of interactions of metal cations with peptides. As an example, the conformational search approach applied for finding the global minima of the gas-phase systems  $\text{AcPheAla}_5\text{LysH}^+$  and  $\text{AcPheAla}_6 + \text{Na}^+$  in Chapter 3 relied on the usage of conventional force fields (FFs) and different levels of DFA. In order to select a rather small number of conformers out of the large pool of structures obtained after sampling the whole conformational space using conventional FFs, a simple energy criterion was applied not only using the calculated FF energies but also energies at the DFA level of GGA xc functionals. This was done because the reliability of FFs for quantitative predictions for systems different from those they were trained on is anything but clear and, in fact, can be misleading, as explained in Subsection 2.3.1. Furthermore, the confident assignment of calculated IR spectra to their measured counterparts on top of an accurate energy hierarchy finding in the low-energy region required the usage of computationally costly hybrid xc functionals. These are but two examples that motivated the research presented here whose goal is to investigate the energetics of peptides in conjunction with metal cations. That is to assess the goodness of commonly applied theoretical levels of theory, *i.e.* FFs, semi-empirical quantum chemistry methods, DFAs, and wavefunction-based methods by evaluating them with respect to high-level coupled-cluster calculations. The focus thereby lies on benchmark systems in the gas phase consisting of either a bare acetylhistidine (AcH) or microsolvated with a  $\text{Zn}^{2+}$  cation. Besides the examples of metalloproteomics given in the beginning of Chapter 1, the choice for the system of AcH has been made because it is still computationally feasible, even for high-level methods, yet provides a challenging structure because of the tautomeric form of its neutral imidazole ring that has already been depicted in Figure 2.2 in Section 2.1.

Figure 4.1 shows chemical structures of AcH with the different protonation states investigated in this work: Negatively charged AcH (upper row in Figure 4.1) has two equivalent tautomeric forms of the neutral imidazole side chain. The two forms are labeled  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^-$  and  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^-$ , meaning that either the  $\text{N}_{\delta_1}$  or the  $\text{N}_{\epsilon_2}$  atom is protonated in the imidazole ring. For bare neutral AcH (bottom row in Figure 4.1), three different protonation states are theoretically possible: Besides the two equivalent tautomeric forms, labeled  $\text{AcH}(\text{N}_{\delta_1})-\text{COOH}$  and  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH}$ , that have a neutral carboxyl group at the C-terminus ( $-\text{COOH}$ ), a third form exists, labeled  $\text{AcH}^+-\text{COO}^-$ , which has both the  $\text{N}_{\delta_1}$  and  $\text{N}_{\epsilon_2}$  nitrogens of the imidazole protonated but the carboxyl group at the C-terminus deprotonated ( $-\text{COO}^-$ ). As already pointed out, either system is studied bare as well as microsolvated with a  $\text{Zn}^{2+}$  cation, resulting

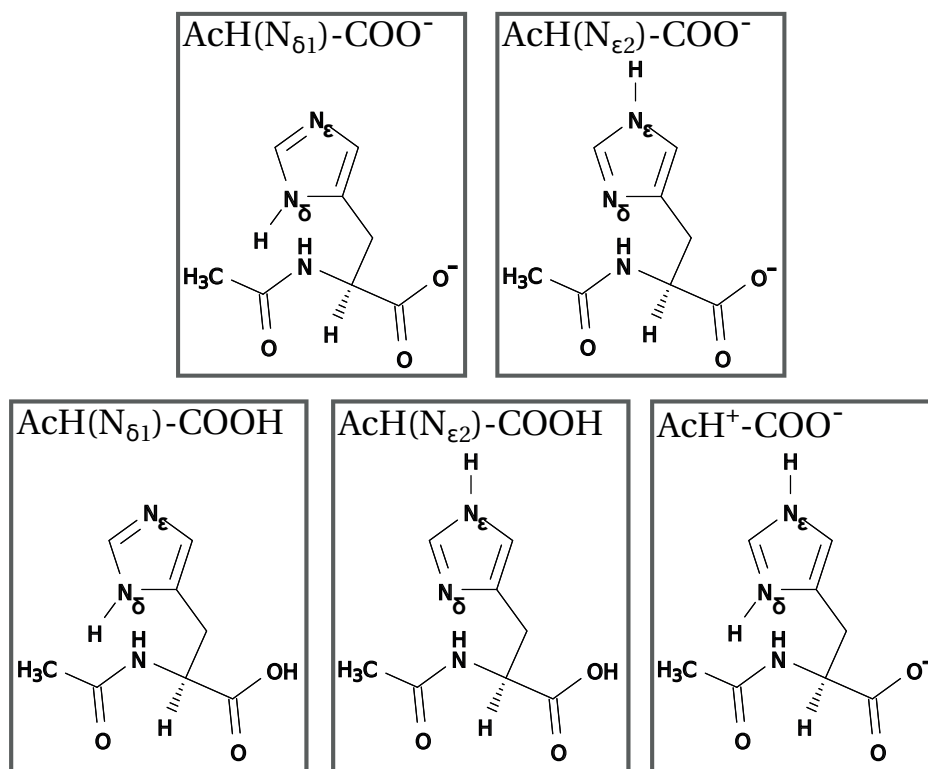


Figure 4.1 – Chemical structures of negatively charged AcH (upper row) showing the two equivalent tautomeric forms of the neutral imidazole side chain. For neutral AcH (bottom row), three different protonation states are theoretically possible.

in ten different systems to be investigated.

Various benchmark calculations for small systems containing a zinc cation have been done in the past. Amin and Truhlar set up a benchmark database of Zn coordination compounds with O, S,  $\text{NH}_3$ ,  $\text{H}_2\text{O}$ , OH,  $\text{SCH}_3$ , and H ligands [350]. Using coupled cluster calculations with augmented polarized triple-zeta basis sets as the reference, 39 density functionals and seven more approximate molecular orbital theories were tested. They found that DFT overall significantly outperformed semi-empirical methods. Best performance was generally found for xc functionals containing a portion of Hartree-Fock exchange, *i.e.* hybrid functionals. Out of the functionals that contained no Hartree-Fock exchange, M06-L (see Subsection 2.3.5) displayed the best performance. Similarly, Rayón *et al.* tested the performance of five different functionals against MP2 (see Subsection 2.3.7) and CCSD(T) (see Subsection 2.3.8) calculations, with the B3LYP functional performing best [351]. Weaver *et al.* predicted nine  $\text{ZnX}$  complexes ( $\text{X} = \text{Zn}, \text{H}, \text{O}, \text{F}_2, \text{S}, \text{Cl}, \text{Cl}_2, \text{CH}_3, (\text{CH}_3)_2$ ) using 14 density functionals, MP2 calculations and the CCSD and CCSD(T) coupled-cluster methods applying correlation consistent triple-zeta basis sets [352]. Comparing heats of formation against experimentally determined values, they found that BLYP, B3LYP, MP2, CCSD and CCSD(T) showed poor performances based on accuracy, which for the latter three wavefunction based methods might be caused by a miss-

ing complete basis set description (see Subsection 2.3.9) or the slow-converging correlation contribution of the zinc electrons that may lead to large and conformation dependent basis set superposition errors (BSSE). Gutten *et al.* evaluated the performance of the wavefunction-based MP2 method as well as several DFA xc functionals with respect to CCSD(T) using gas-phase complexation energies calculated for five model complexes and four metal ions (Fe<sup>2+</sup>, Cu<sup>2+</sup>, Zn<sup>2+</sup>, Cd<sup>2+</sup>) [353]. Reasonable agreement was found for MP2 with values usually within 1.5 kcal/mol from the reference values, while DFT performed less satisfactory, although the appropriateness of the models may be significantly altered when combining them with advanced solvation models [354]. For certain complexes containing metal-ligand bonds, large errors in the gas-phase complexation energies (with values up to 20 kcal/mol) were reported. Performance concerning geometry optimization was found to be satisfactory already using the PBE xc functional on the GGA DFT level. In the benchmark studies by Navrátil *et al.* on activation and reaction energies for four model systems of peptide bond hydrolysis in an ion-free environment and in presence of one and two zinc ions, reasonably good performance was found for several DFAs and MP2 when comparing to CCSD(T)-obtained results [355]. Best performance for calculating activation barriers was achieved when using the B3LYP or the M06-2X xc functionals on the DFA level of theory. Finally, benchmark evaluations and calibrations of theoretical calculations help in modeling metal-binding sites and studying metal-ion selectivity in proteins [356–359].

The general approach followed here is briefly outlined in the following. First, a global search for minima on the PES combining both FF and DFA is performed for either one of the ten systems individually. The obtained global minima and energy hierarchies are then discussed and compared for systems of equal overall charge  $q$ , *i.e.*  $q = 1$  for the upper row in Figure 4.1 and  $q = 0$  for the bottom row in Figure 4.1. For the benchmarking studies, a set of structures is then selected based on simple energy criteria. While the focus does lie on local minima structures, it is intended to select structures that vary in energy and structure in order to intentionally provide a challenge for the theoretical methods to be benchmarked. On top of that, all systems carrying the same overall charge  $q$  are benchmarked at once (except for FFs), thus providing even more challenge for the methods in question. Finally, across-the-scale total energy calculations for a wide variety of FFs, semi-empirical methods, DFAs, and wavefunction based methods are tested and evaluated against high-level coupled cluster calculations using mean absolute errors (MAEs) and maximum errors (MEs) as a quality measure, as explained in Subsection 4.2.3.

## 4.2 Computational Details

### 4.2.1 Conformational Sampling

In order to yield minima structures that serve as a basis for selecting a set of conformers for the benchmarking process, the conformational space needs to be sampled first. To that end, an energy minimum search combining both FF and DFA very similar to the one used in

Section 3.3 is laid out. First, a global energy minimum search is performed using the basin-hopping approach within the TINKER molecular modeling package, as explained in detail in Section 2.4. Here, the 2009 AMOEBA biopolymer force field, labeled AMOEBA-BIO09 (see Subsection 2.3.1), is applied, which is for two reasons: First, this polarizable force field provides a much “rougher” potential energy surface than widely used conventional force fields, such as AMBER-99, CHARMM22, or OPLS-AA, because it uses atomic charge multipole expansion instead of fixed point charges. The “rougher” the potential energy surface the more minima are found, hence the conformational space is sampled in more detail. For example, depending on the actual studied system, *i.e.* whether the  $\text{Zn}^{2+}$  cation is present or the protonation state of the imidazole side chain and the carboxyl group, the number of minima found can be up to a factor of six higher when using the AMOEBA-BIO09 force field in comparison to the OPLS-AA, AMBER-99, and CHARMM22 force fields. Secondly, the AMOEBA-BIO09 FF is the only FF available providing out-of-the-box parameters for the neutral carboxyl group ( $-\text{COOH}$ ). Concerning the technical aspect of the basin-hopping search, the `scan` subprogram within TINKER has been applied using all automatically found torsional angles, a relative energy window of 100kcal/mol and an energy similarity criterion of 0.0001 kcal/mol. After having applied the FF driven basin-hopping approach, all found minima are locally refined using DFA implemented within FHI-aims. Like described in Section 3.3, local refinement is done first on the PBE+vdW<sup>TS</sup> level using FHI-aims specific `tier 1` basis sets and `light` settings intended to give reliable energies for screening purposes [264]. After clustering, further relaxation is accomplished at the PBE+vdW<sup>TS</sup> level using `tier 2` basis sets and `tight` settings that are intended to provide meV-level accurate energy differences [264], *i.e.* within 0.02 kcal/mol. Finally, relaxation is accomplished at the PBE0+MBD level using the same two-step approach as before, *i.e.* using first `tier 1` basis sets and `light` settings, and `tier 2` basis sets and `tight` settings afterwards.

### 4.2.2 Levels of Theory and Energy Calculation Methods

Applied energy calculation methods and levels of theory have been discussed in great detail in Section 2.3 and are thus only briefly summarized in the following, including necessary technical details for calculation.

The benchmark calculations are based on high-level coupled-cluster calculations (see Subsection 2.3.8). In particular, the coupled-cluster method including single, double, and perturbative triple excitations, named CCSD(T), is commonly referred to as the “gold standard of quantum chemistry” due to its high accuracy in the complete basis set limit (CBS) [251, 252]. However, due to the slow convergence of the electronic correlation energy with basis set size  $N$  as well as the technique’s  $\mathcal{O}(N^7)$ -scaling of the computational costs, accurate results that require large enough basis sets are currently not affordable for system sizes treated in this work. Instead, the domain-based local pair natural orbital (DLPNO-)CCSD(T) technique serves as the reference method in this work. As laid out in Subsection 2.3.8, the DLPNO-CCSD(T) approximation aims to fully exploit locality of the electron correlation and shows a near linear

scaling behavior with basis set size  $N$ . Calculations are carried out with ORCA, while Ahlrichs' def2 basis set family (see Subsection 2.3.9) is used for all wavefunction-based methods. Because heavy elements like Zn<sup>2+</sup> require a relativistic treatment of all-electron calculations, the 0<sup>th</sup> order regular approximation (ZORA) (see Subsection 2.3.9), implemented in ORCA in an approximate way [272, 273], is used throughout. As the scalar relativistic treatment requires flexible basis sets, this in turn means that ORCA automatically provides relativistically recontracted versions [274] of Ahlrichs' def2 basis set family, labeled ZORA-def2. The accuracy of the DLPNO-CCSD(T) method has been tested previously with a series of benchmark sets covering a broad range of quantum chemical applications [360]. An accuracy of 1 kcal/mol commonly referred to as "chemical accuracy", could be obtained using normal settings. Still, before using the DLPNO-CCSD(T) method with normal settings as the reference method in this work, validation has to be done against conventional CCSD(T) calculations for the systems depicted in Figure 4.1 and using Ahlrichs' relativistically recontracted split valence basis set with added polarization functions, labeled ZORA-def2-SVP.

The other post-Hartree-Fock *ab initio* method in this work to be benchmarked against DLPNO-CCSD(T) is the widely used second-order Møller-Plesset perturbation theory (MP2) (see Subsection 2.3.7). Calculations are carried out again with ORCA and applying a resolution of identity (RI) approximation [361].

Energy calculations for both DLPNO-CCSD(T) and MP2 are performed using Ahlrichs' ZORA-def2-SVP basis set as well as relativistically recontracted valence triple-zeta and quadruple-zeta basis sets with two sets of polarization functions added, labeled ZORA-def2-TZVPP and ZORA-def2-QZVPP, respectively. Extrapolation to the CBS limit is applied on calculated Hartree-Fock (HF) energies and correlation energies individually, as laid out in detail in Subsection 2.3.9. HF energies are extrapolated using the form proposed by Karton and Martin given in Equation (2.156), while the extrapolation scheme for the correlation energies follows the form proposed by Truhlar given in Equation (2.157). Extrapolation using all three basis set families has been found to yield inconsistent results between the different systems depicted in Figure 4.1. Hence, extrapolation is laid out using only ZORA-def2-TZVPP and ZORA-def2-QZVPP, resulting in an effective two-point extrapolation scheme using  $n = 3, 4$  and assuming  $\beta = 3$  in Equation (2.157), as originally proposed by Halkier *et al.* [277].

Finally, for systems microsolvated with a Zn<sup>2+</sup> cation, the slow-converging correlation contribution of the zinc electrons may lead to large and conformation dependent basis set superposition errors (BSSE). To account for that and prior to performing CBS extrapolation, the HF and correlation energies of each Zn<sup>2+</sup> coordinated conformation are subjected to the counterpoise correction as proposed by Boys and Bernardi assuming rigid conformers: Following Equation (2.158), the BSSE is estimated as

$$\begin{aligned} E_{\text{BSSE}} &= E_{\text{BSSE}}(\text{AcH}) + E_{\text{BSSE}}(\text{Zn}^{2+}), \\ \text{with } E_{\text{BSSE}}(\text{AcH}) &= E^{\text{AcH}+\text{Zn}^{2+}}(\text{AcH}) - E^{\text{AcH}}(\text{AcH}), \\ \text{and } E_{\text{BSSE}}(\text{Zn}^{2+}) &= E^{\text{AcH}+\text{Zn}^{2+}}(\text{Zn}^{2+}) - E^{\text{Zn}^{2+}}(\text{Zn}^{2+}), \end{aligned} \quad (4.1)$$

where  $E^{\text{AcH}+\text{Zn}^{2+}}(\text{AcH})$  represents the energy of AcH evaluated in the union of the basis functions associated with AcH and  $\text{Zn}^{2+}$ ,  $E^{\text{AcH}}(\text{AcH})$  represents the energy of AcH evaluated in the basis functions associated with AcH, *etc.* The individual BSSEs are then subtracted from the Hartree-Fock and correlation energy, respectively.

Single-point energy calculations using several out-of-the-box force fields (FFs) are carried out using the TINKER molecular modeling package. Two classes of FFs are tackled: (i) conventional FFs, in particular AMBER-99, CHARMM22, and OPLS-AA, as well as (ii) polarizable atomic multipole-based FFs that use atomic charge multipole expansion instead of fixed point charges. In particular, these are the 2009 AMOEBA biopolymer FF named AMOEBA-BIO09, and the 2013 AMOEBA protein FF named AMOEBA-PRO13. A detailed description of the different FFs is provided in Subsection 2.3.1. Because of the intrinsic concept of FFs that requires *a priori* definition of bonds, angles, torsions, *etc.* along with the corresponding parameters, different protonation states are not comparable in energy. Hence, energies of conformers may only be benchmarked if the structures correspond to the same protonation state. Note that only for systems containing a deprotonated carboxyl group ( $-\text{COO}^-$ ), parameters are available for all force fields out-of-the-box. As AMOEBA-BIO09 is the only FF available providing also parameters for the neutral carboxyl group ( $-\text{COOH}$ ), FF calculations for systems containing neutral AcH (lower row in Figure 4.1) are only laid out using this particular FF.

Semi-empirical quantum chemistry methods are based on the Hartree-Fock method, but follow a simplification strategy by making approximations for computationally demanding terms. In order to account for caused errors, empirical parameters are incorporated into the formalism and fitted against experimental data or high-level calculations [128]. Details are provided in Subsection 2.3.4. All semi-empirical methods tackled in this work are based on the neglect of diatomic differential overlap (NDDO), a method for approximating computational costly three-center and four-center two-electron integrals, as laid out in Subsection 2.3.4. In particular, the different applied models are the Austin Model 1 (AM1), the Parametric Method 3 (PM3), the Parametric Method 6 (PM6), and the Parametric Method 7 (PM7). All semi-empirical method calculations have been carried out using the MOPAC2016 [261] semi-empirical quantum chemistry program. For the specific case of PM6, two additional long-range dispersion correction schemes are tackled as well. In particular, these are Grimme's D3 correction for dispersion plus a simple function for hydrogen bonds, as well as the corrections to hydrogen bonding and dispersion by Řezáč and Hobza, labeled D3H4. The corresponding conjunctive methods are then accordingly being labeled PM6-D3 and PM6-D3H4.

As explained in Subsection 2.3.4, semi-empirical energy evaluations yield heats of formation as the respective semi-empirical methods are parameterized on experimental heats of formation [262]. That is in contrast to the other methods tackled in this work for which energy calculations refer to total energies on the PES. However, when comparing potential energies of other computational methods with heats of formation obtained from semi-empirical calculations through the means of MAEs and MEs, the systematic shift between the two is accounted for, as explained in Subsection 4.2.3.

## Chapter 4. Energetics and Benchmark of Energy Methods of Acetyl-Histidine with Zn<sup>2+</sup>

---

Concerning DFT in itself which is an exact method, in practice approximations have to be made because the exact form of the xc functional is unknown, except for the free electron gas. As laid out in detail in Subsection 2.3.5, a large variety of different DFAs exist, commonly classified into different types depending on the features and formal properties of the xc functionals in question. The ones selected in this work are summarized in the following:

- *Generalized gradient approximations (GGAs)* are characterized by the dependence of the xc functional only on the electron density and its gradient. In this work, the accuracy of the Perdew-Burke-Ernzerhof (PBE) and Becke-Lee-Yang-Parr (BLYP) xc functionals is studied.
- In addition to GGAs, *meta*-GGAs also depend on the Laplacian of the electron density or include the kinetic energy density. Here, the M06-L and M11-L xc functionals from the group of Minnesota functionals are tested as well as the SCAN functional.
- For the computationally more costly class of *hybrid* functionals, the exchange parts of the functional are admixed with exact exchange from Hartree-Fock theory. Here, the PBE0, B3LYP, and SCAN0 functionals are tested. In addition, several hybrid functionals from the group of Minnesota functionals are tackled as well, in particular the M06, M06-2X, M08-SO, M08-HX, and M11 functionals.

Calculations for the PBE, BLYP, M11-L, SCAN, PBE0, B3LYP, M08-SO, M08-HX, and M11 xc functionals are carried out with FHI-aims using `tier 2` basis sets and `really_tight` settings, and including a relativistic treatment by applying the atomic ZORA method. The SCAN and SCAN0 functionals are implemented in FHI-aims via the `default` program [362]. Calculations for the M06-L, M06, and M06-2X xc functionals are carried out with ORCA, including ZORA and the relativistically recontracted ZORA-`def2-QZVPP` basis set, as explained above.

Commonly applied semi-local DFAs and conventional hybrid functionals are unable to capture the essence of long-range dispersion effects. As laid out in detail in Subsection 2.3.6, many systems containing biomolecules rely on vdW interaction treatments for an accurate energetic description. Three different *a posteriori* vdW correction schemes are tackled in this work:

- The general empirical additive D3 dispersion correction method by Grimme *et al.* provides a consistent description across the whole periodic table. Here, the zero-damping function for short ranges is used, including three-body dispersion contributions. In order to match the long- and midrange correlation of D3 with the semilocal correlation computed by the xc functional, the parameterization of the damping function depends on the xc functional itself. Hence, only xc functionals where an out-of-the-box D3 treatment is available are tested. In particular, M06-L+D3, M06+D3, and M06-2X+D3 are evaluated using ORCA and applying the same settings as described above. For the methods of PBE+D3, BLYP+D3, PBE0+D3, and B3LYP+D3, long-range dispersion calculations are done on top of the FHI-aims calculated energies using Grimme's stand-alone program DFT-D3 [363].



- The parameter-free pairwise Tkatchenko-Scheffler van der Waals scheme (vdW<sup>TS</sup>) relies on summing interatomic pairwise, electron-density derived  $C_6$  coefficients, and accurate reference data for the free atoms. As the method is implemented in FHI-aims, calculations are carried out for the methods of PBE+vdW<sup>TS</sup>, BLYP+vdW<sup>TS</sup>, PBE0+vdW<sup>TS</sup>, and B3LYP+vdW<sup>TS</sup>.
- In contrast to the previous pairwise Tkatchenko-Scheffler scheme that ignores the intrinsic many-body nature of correlation effects, the many-body dispersion scheme labeled MBD (and sometimes also labeled MBD\* or MBD@rsSCS) combines the TS scheme with the self-consistent screening (SCS) equation of classical electrodynamics. In addition, a range-separation (rs) technique is applied, separating correlation into a short-range and a long-range contribution. Details are provided in Subsection 2.3.6. Calculations are carried out for the methods of PBE+MBD and PBE0+MBD using FHI-aims.

In order to avoid high computational costs of hybrid xc functionals and still yield accurate results, recent focus has been set on “low-cost” DFT based *composite electronic structure approaches*. In particular, the PBEh-3c method by Grimme *et al.* aims to efficiently compute structures and interaction energies, as laid out in detail in Subsection 2.3.5. Calculations are carried out with ORCA.

Finally, *double hybrid xc functionals* extend hybrid xc functionals in a way that both the exchange and the correlation part contain non-local orbital-dependent components, as explained in detail in Subsection 2.3.5. In particular, the B3LYP+XYG3 method is tested. Calculations are carried out with FHI-aims using numerically tabulated atom-centered orbital triple-zeta basis sets with valence-correlation consistency, labeled NA0-VCC-nZ [267] (see Subsection 2.3.9). Zhang *et al.* showed that XYG3 provides best results in combination with the triple-zeta NA0-VCC-3Z basis set [270]. Because the NA0-VCC-3Z basis set is not available out-of-the-box for the element of Zn, Dunning’s analogous cc-pVTZ [269] basis set is used instead for this particular element.

#### 4.2.3 Mean Absolute Error (MAE) and Maximum Error (ME)

In order to compare the energetic performance of different methods, single-point energy calculations of a set of different conformers are compared by means of mean absolute errors (MAEs) and maximum errors (MEs). MAEs of relative energies between the reference method and the method to be benchmarked are calculated as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\Delta E_i^{\text{reference}} - \Delta E_i^{\text{benchmarked}} + c|, \quad (4.2)$$

where the index  $i$  runs over all  $N$  conformations of a given data set.  $\Delta E_i$  in principle denotes the energy difference between conformer  $i$  and the lowest-energy conformer of the set. The adjustable parameter  $c$  is used to systematically shift the reference and benchmark confor-

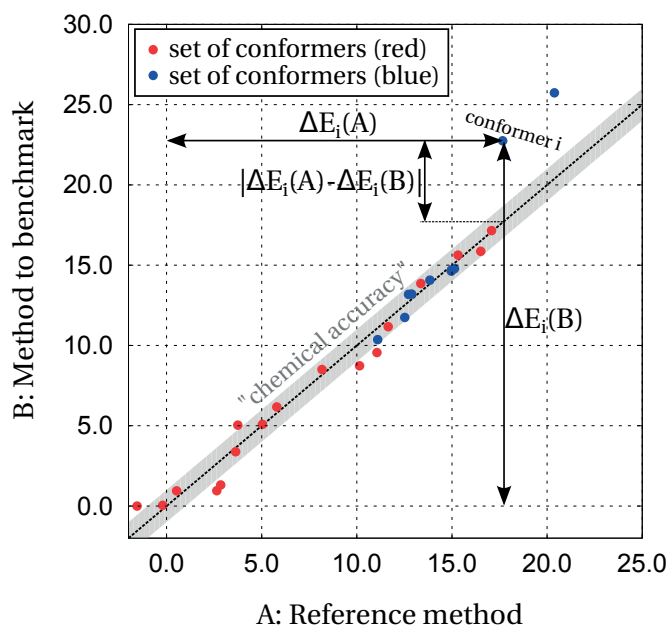


Figure 4.2 – Example of a correlation plot of two different sets of conformers (red and blue). The reference (A) and benchmark (B) conformational hierarchies have already been shifted uniformly to minimize the MAE. If the energy description between the reference method and the method to be evaluated agreed perfectly, all points would align on the dashed diagonal line. The gray shading denotes a corridor of an absolute energy deviation of 1 kcal/mol, *i.e.* the region of “chemical accuracy”. For a specific conformer  $i$ , the absolute energy deviation  $|\Delta E_i^{\text{reference}} - \Delta E_i^{\text{benchmarked}}| = |\Delta E_i(A) - \Delta E_i(B)|$  is illustrated.

mational hierarchies versus one another to obtain the lowest possible MAE, rendering the reported MAE value independent of the choice of any reference structure. Similarly, MEs are calculated as follows:

$$\text{ME} = \max_{i \in N} |\Delta E_i^{\text{reference}} - \Delta E_i^{\text{benchmarked}}| + c, \quad (4.3)$$

using the same notation as above. Figure 4.2 shows an example of a correlation plot including a graphical illustration of  $|\Delta E_i^{\text{reference}} - \Delta E_i^{\text{benchmarked}}|$ .

## 4.3 Results

### 4.3.1 Energy Hierarchies

Figure 4.3 shows the obtained energy hierarchies at the PBE0+MBD level after having completed the conformational search for each individual protonation state of bare negatively charged AcH and bare neutral AcH, as well as both systems in presence of a  $\text{Zn}^{2+}$  cation. Figure 4.4 illustrates the structure of the lowest-energy conformer for each depicted protonation state.

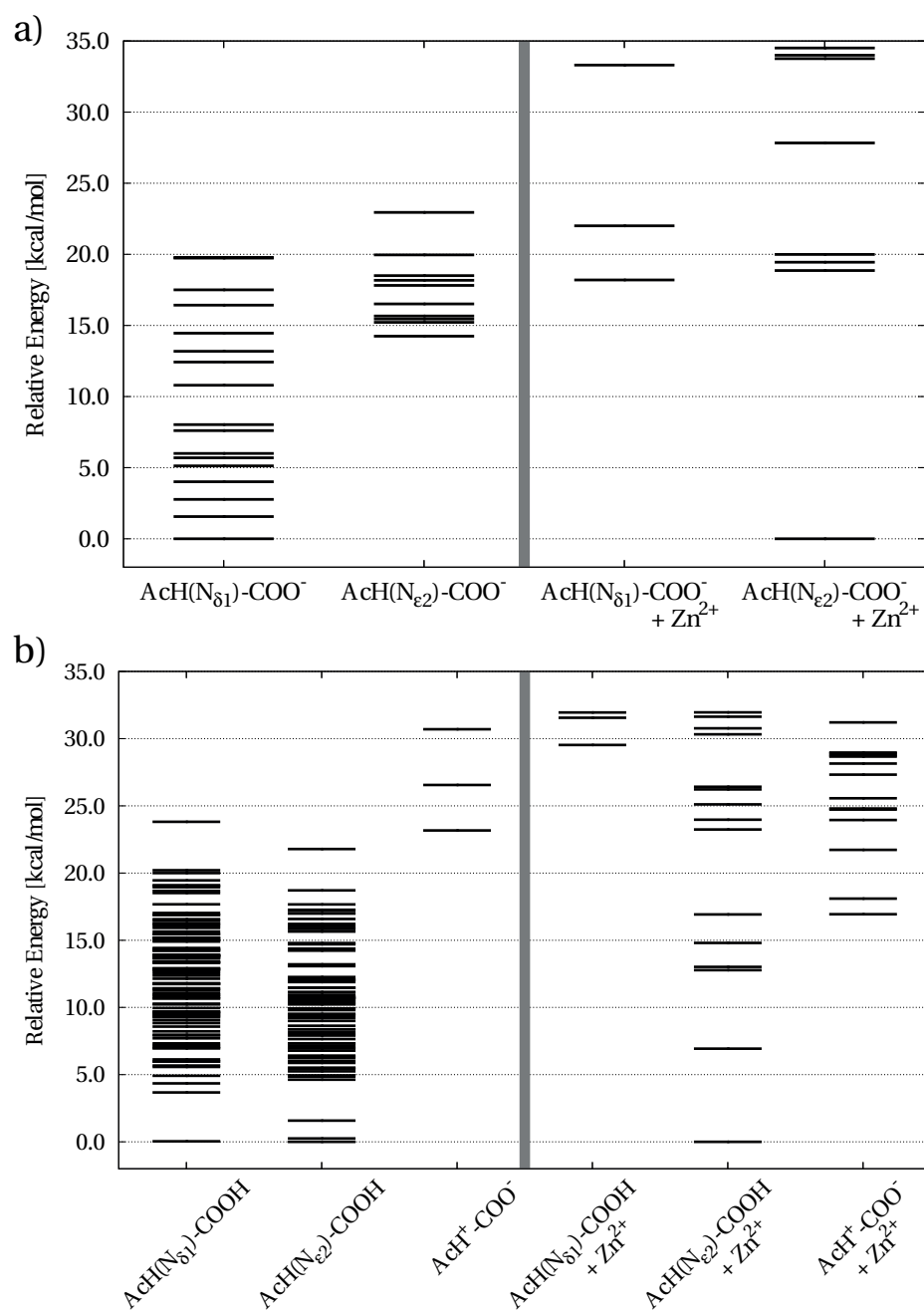


Figure 4.3 – Obtained energy hierarchies at the PBE0+MBD level after having completed the conformational search for (a) negatively charged AcH, bare and with an additional  $\text{Zn}^{2+}$ , and (b) neutral AcH, bare and with an additional  $\text{Zn}^{2+}$ .

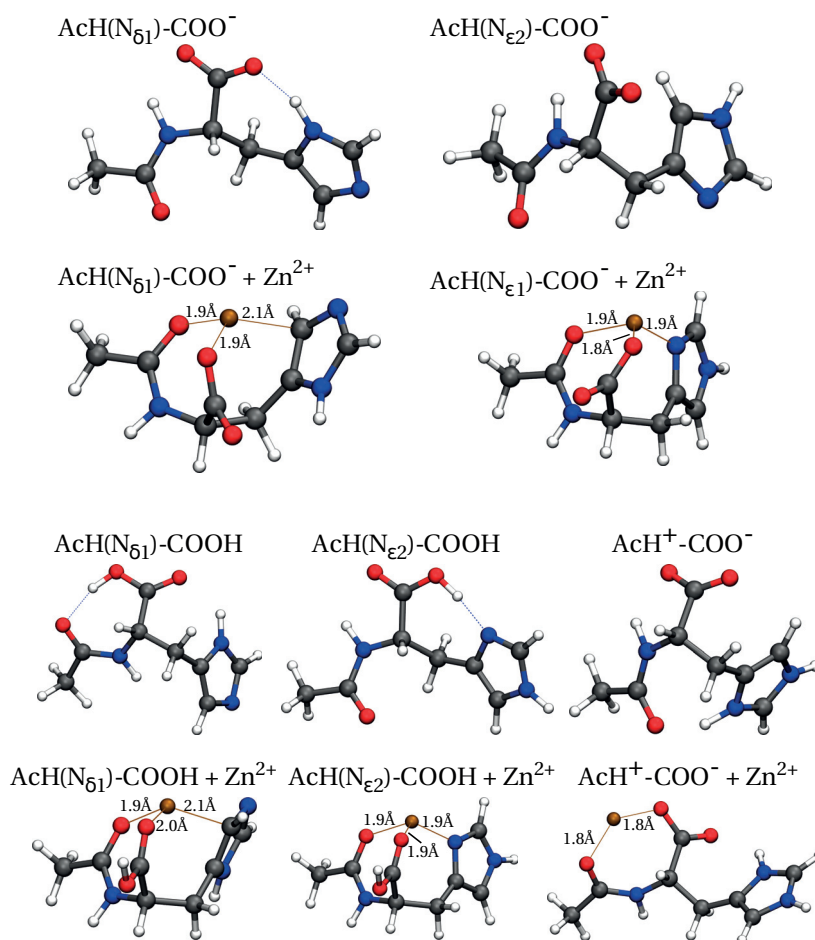


Figure 4.4 – Illustration of the structure of the lowest-energy conformer for each depicted protonation state.

Comparing the two possible protonation states for negatively charged AcH, *i.e.*  $\text{AcH}(\text{N}_{\delta 1})-\text{COO}^-$  and  $\text{AcH}(\text{N}_{\epsilon 2})-\text{COO}^-$  (see Figure 4.3(a)), it is immediately evident that the protonation of the nitrogen atoms of the imidazole ring has a large impact concerning energy and structure of the system. The lowest-energy conformer of  $\text{AcH}(\text{N}_{\delta 1})-\text{COO}^-$  lies 14.3 kcal/mol lower in energy than the lowest-energy conformer of  $\text{AcH}(\text{N}_{\epsilon 2})-\text{COO}^-$ , meaning that the tautomeric state of having the  $\text{N}_{\delta 1}$  nitrogen atom of the imidazole ring protonated is energetically favored over having the proton residing at the  $\text{N}_{\epsilon 2}$  nitrogen atom. The reason for that comes abundantly clear when comparing the two lowest-energy conformers that are illustrated in Figure 4.4: In the case of  $\text{AcH}(\text{N}_{\delta 1})-\text{COO}^-$ , there exists the geometrical possibility of forming a hydrogen bond between one oxygen of the anionic carboxylate group at the C-terminus and the nitrogen-bound hydrogen. In case of having the  $\text{N}_{\epsilon 2}$  nitrogen atom protonated, a hydrogen bond cannot be formed as the proton “points away” from the anionic carboxylate group, explaining the much higher energy of this structure in comparison with its tautomeric counterpart.

The situation however changes drastically when introducing a  $\text{Zn}^{2+}$  cation to the system. As seen in Figure 4.3(a), the lowest-energy conformer of  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$  is now 18.2 kcal/mol higher in energy than the lowest-energy conformer of  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$ . The corresponding structures illustrated in Figure 4.4 look fairly similar in part as the oxygen atom of the carbonyl group at the acetylated N-terminus as well as one oxygen of the anionic carboxylate group are coordinated towards the  $\text{Zn}^{2+}$ . They differ however in the different orientation of the imidazole ring towards the cation. In the case of  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$ , the deprotonated  $\text{N}_{\delta_1}$  atom allows for a coordinate bonding interaction with the  $\text{Zn}^{2+}$  cation, resulting in an energetically more favorable structure compared to  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$  where the deprotonated  $\text{N}_{\epsilon_2}$  atom points away from the cation, resulting in an energetically less favorable cation- $\pi$  interaction between imidazole ring and cation. Adding a  $\text{Zn}^{2+}$  cation to the system also results in an increased energetic gap between conformers. For example, the two lowest-energy conformers of  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^-$  are separated by 1.6 kcal/mol while the gap increases to 3.8 kcal/mol for  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$ . For  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^-$ , the two lowest-energy conformers are separated by 1.0 kcal/mol, while the gap increases to 18.9 kcal/mol for  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$ .

The hierarchies of the three different protonation states of bare neutral AcH are shown in Figure 4.3(b). The global-minimum conformers of the systems of  $\text{AcH}(\text{N}_{\delta_1})-\text{COOH}$  and  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH}$  are very similar in energy, differing only by 0.04 kcal/mol. For  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH}$ , a hydrogen bond is possible between the deprotonated  $\text{N}_{\delta_1}$  atom and said proton, resulting in a very similar structure compared to system  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^-$ , as shown in Figure 4.4. For  $\text{AcH}(\text{N}_{\delta_1})-\text{COOH}$ , due to the protonated  $\text{N}_{\delta_1}$  atom, the proton at the carboxyl group points away from the imidazole ring and is coordinated towards the N-terminus, forming a hydrogen bond with the carbonyl group. A protonated imidazole ring, as seen in the protonation state of system  $\text{AcH}^+-\text{COO}^-$ , results in an energetically unfavorable structure, being 23.2 kcal/mol higher in energy than the global minimum of system  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH}$ .

The situation changes again when introducing a  $\text{Zn}^{2+}$  cation to the system. The system of  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH} + \text{Zn}^{2+}$  is energetically most favorable as the structure of the global minimum is very similar to the one of the system of  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$ : The deprotonated  $\text{N}_{\delta_1}$  atom allows for a coordinate bonding interaction with the  $\text{Zn}^{2+}$  cation that in turn is also coordinated towards the electronegative oxygen atoms at the carboxyl group at the C-terminus and the carbonyl group at the N-terminus. The global minimum of  $\text{AcH}^+-\text{COO}^- + \text{Zn}^{2+}$  is 16.9 kcal/mol higher in energy than the global minimum of  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH} + \text{Zn}^{2+}$ . The positively charged cation and the protonated imidazole ring share no proximity, resulting in a lowest-energy structure where the  $\text{Zn}^{2+}$  is coordinated between the oxygen of the carbonyl group at the N-terminus and one oxygen of the carboxyl group at the C-terminus. The structure of the global minimum for  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$  is very similar to the one for  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$ , save the twisted imidazole ring due to the protonated  $\text{N}_{\delta_1}$  atom. Similarly to system  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$ , this results in an energetically less favorable cation- $\pi$  interaction between imidazole ring and cation as the global minimum is 29.5 kcal/mol higher in energy than the global minimum of  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$ .

## Chapter 4. Energetics and Benchmark of Energy Methods of Acetyl-Histidine with $\text{Zn}^{2+}$

Table 4.1 – Minima selection criteria across the tackled systems and protonation states.

system	total system charge	energy cut-off	# minima
AcH( $\text{N}_{\delta 1}$ )-COO <sup>-</sup>	-1	23.0 kcal/mol	17*
AcH( $\text{N}_{\epsilon 2}$ )-COO <sup>-</sup>			10*
AcH( $\text{N}_{\delta 1}$ )-COO <sup>-</sup> + Zn <sup>2+</sup>	+1	41.5 kcal/mol	9
AcH( $\text{N}_{\epsilon 2}$ )-COO <sup>-</sup> + Zn <sup>2+</sup>			9
AcH( $\text{N}_{\delta 1}$ )-COOH	0	7.0 kcal/mol	11
AcH( $\text{N}_{\epsilon 2}$ )-COOH			18
AcH <sup>+</sup> -COO <sup>-</sup>		50.0 kcal/mol	8*
AcH( $\text{N}_{\delta 1}$ )-COOH + Zn <sup>2+</sup>	+2	46.0 kcal/mol	9
AcH( $\text{N}_{\epsilon 2}$ )-COOH + Zn <sup>2+</sup>			18
AcH <sup>+</sup> -COO <sup>-</sup> + Zn <sup>2+</sup>			22

The “energy cut-off” means the relative energy with respect to the global minimum for a given total system charge (*i.e.* taking into account all possible protonation states, compare with Figure 4.3), within which all found minima are taken into account. The last column denotes the number of minima used for benchmarking. Numbers denoted with an asterisk (\*) mean all found minima for this particular protonation state are considered.

### 4.3.2 Selection of Minima Structures

For every protonation state, the lowest-energy structures from the previous global minimum search are selected based on energy criteria. For one, this ensures an emphasis on the most likely structures also seen in experiment as there will always be a bias towards structures with low energy, ignoring individual set-ups or experimental conditions. However, benchmark calculations are done including all possible protonation states for a given overall system charge, except for the case of FFs as explained in Subsection 4.2.2. The large energetic differences between global minima (and consequently other low-energy conformers) of individual protonation states as seen in Figure 4.3 therefore provides a challenging benchmark testing situation for the different methods. Table 4.1 summarizes the different energy selection criteria across the systems and protonation states tackled in this work.

### 4.3.3 Validation of DLPNO-CCSD(T) as the Reference Method

As described in Subsection 4.2.2 and in order to validate DLPNO-CCSD(T) as the reference method used in this work, one needs to check the consistency of the method against conventional CCSD(T), commonly referred to as the “gold standard of quantum chemistry”. Calculations are laid out using Ahlrichs’ relativistically recontracted ZORA-def2-SVP basis set for which CCSD(T) calculations are still affordable with respect to computational costs. Consequently, no extrapolation or counterpoise correction is applied here, as the intent is to compare the “pure” total energy performances of both methods, which – if similar – will justify applying DLPNO-CCSD(T) “instead of” conventional CCSD(T), of course with using larger basis sets, to benchmark the other computational methods. Figures 4.5(a)-(d) show

the corresponding correlation plots for all systems tackled in this work. The alignment of the points near the dashed diagonal line indicates a very similar energy description between the two methods across all systems and protonation states. To quantify that, MAEs and MEs are computed according to Equations (4.2) and (4.3), respectively. For the four different systems, MAEs and MEs are given in Figure 4.5(e). In all cases, MAEs are well within “chemical accuracy”, *i.e.* smaller than 0.5 kcal/mol. Furthermore, MEs are also smaller than 1 kcal/mol for all systems. Taking into consideration that different protonation states and minima that differ in energy by up to more than 50 kcal/mol have been used, one may safely conclude that DLPNO-CCSD(T) serves as a valid reference method for the benchmarking process of other computational methods.

In order to finally yield accurate total energies serving as benchmarks, counterpoise correction needs to be applied following Equation (4.1) and extrapolation to the complete basis set limit is done following Equations (2.156) and (2.157) using Ahlrichs’ relativistically recontracted ZORA-def2-SVP, ZORA-def2-TZVPP, and ZORA-def2-QZVPP basis sets.

#### 4.3.4 Benchmarking Force Fields and Semi-Empirical Methods

Figure 4.6 shows obtained MAEs and MEs calculated according to Equations (4.2) and (4.3) for all systems tackled in this work. As explained in Subsection 4.2.2, FF performance evaluation is treated individually for different protonation states.

Considering bare neutral AcH, see Figure 4.6(a), conventional FFs that make use of fixed point charges are comparable in performance: For AcH(N<sub>δ1</sub>)-COO<sup>-</sup>, MAEs for AMBER-99, CHARMM22, and OPLS-AA have been found to be 2.1 kcal/mol, 2.2 kcal/mol, and 2.4 kcal/mol, respectively. Considering the fact that FF parameters have been derived from systems in solvation instead of gas-phase calculations applied here, the result can be considered satisfactory. However, large MEs with up to 6.9 kcal/mol for OPLS-AA, indicate a possible large deviation in the energetic description for individual conformers. Somehow surprisingly, polarizable atomic multipole-based FFs AMOEBA-BIO09 and AMOEBA-PRO13 perform worse than their FF counterparts using fixed point charges. Large MEs up to 10.9 kcal/mol and 17.3 kcal/mol for AMOEBA-BIO09 and AMOEBA-PRO13, respectively, indicate severe discrepancies in the energetic description for individual conformers. Consequently, the corresponding MAEs of 3.6 kcal/mol and 5.3 kcal/mol are larger than for conventional FFs. Qualitative similar results are found for AcH(N<sub>e2</sub>)-COO<sup>-</sup>. Best performance for FFs is found using CHARMM22 with a MAE of 1.5 kcal/mol and a ME of 3.3 kcal/mol.

Semi-empirical methods show a comparable performance to FFs, but carry the advantage over FFs to be able to describe both protonation states simultaneously. Best performance is found for PM7 with a MAE of 1.7 kcal/mol and a ME of 5.5 kcal/mol. For PM6, adding a long-range dispersion treatment method, *i.e.* D3 or D3H4, yields very similar results of approximately 1.9 kcal/mol, as is expected for a system of such small size.

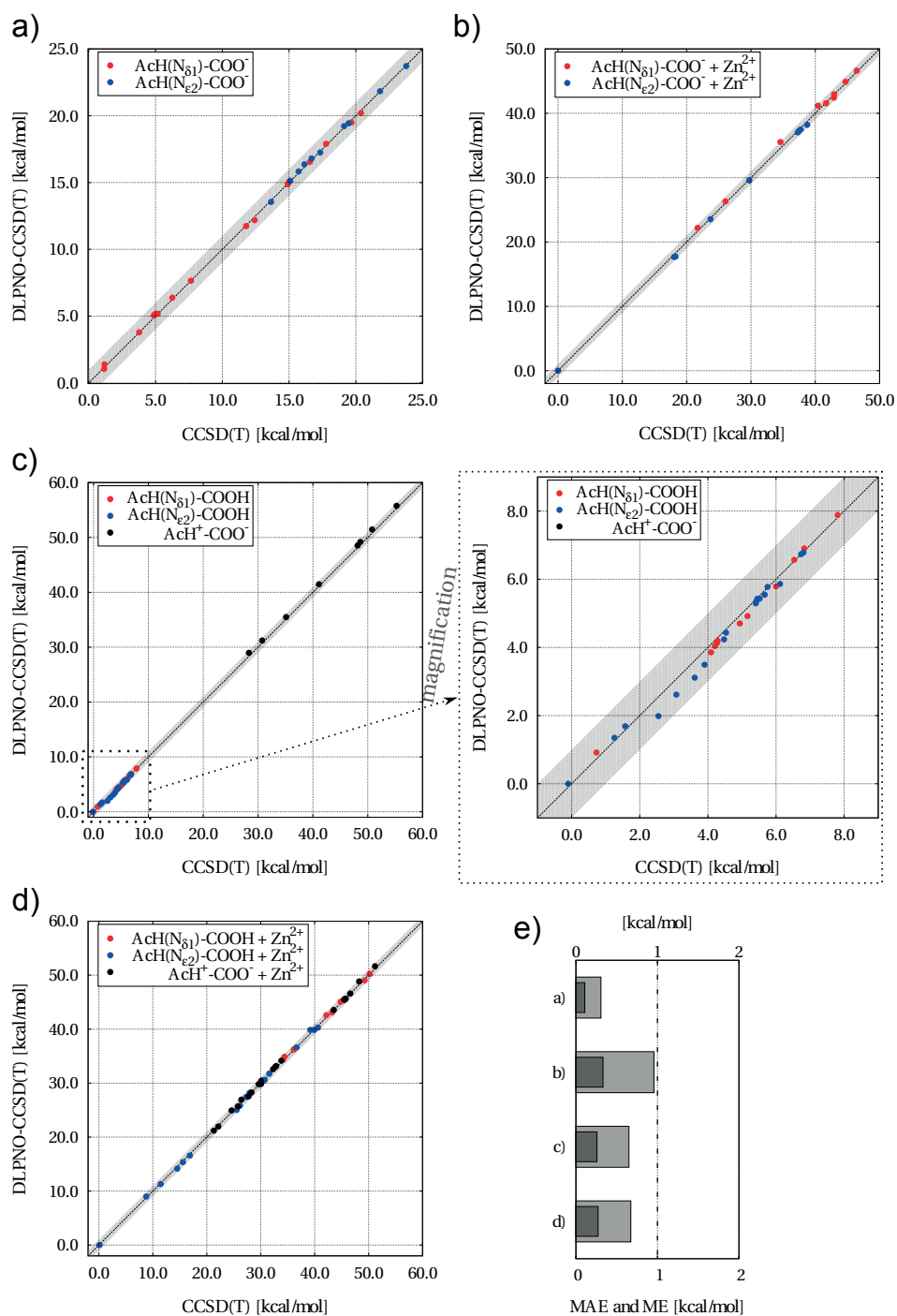


Figure 4.5 – Correlation plots for benchmarking DLPNO-CCSD(T) against conventional CCSD(T) using the ZORA-def2-SVP basis set. The systems tackled refer to (a) negatively charged AcH, (b) the same protonation states in presence of a  $\text{Zn}^{2+}$  cation, (c) bare neutral AcH, and (d) the same protonation states in presence of a  $\text{Zn}^{2+}$  cation. The gray shading denotes an absolute energy deviation of 1 kcal/mol, *i.e.* the region of “chemical accuracy”. (e) Obtained MAEs (dark-gray) and MEs (light-gray) for the four systems, following Equations (4.2) and (4.3), respectively.



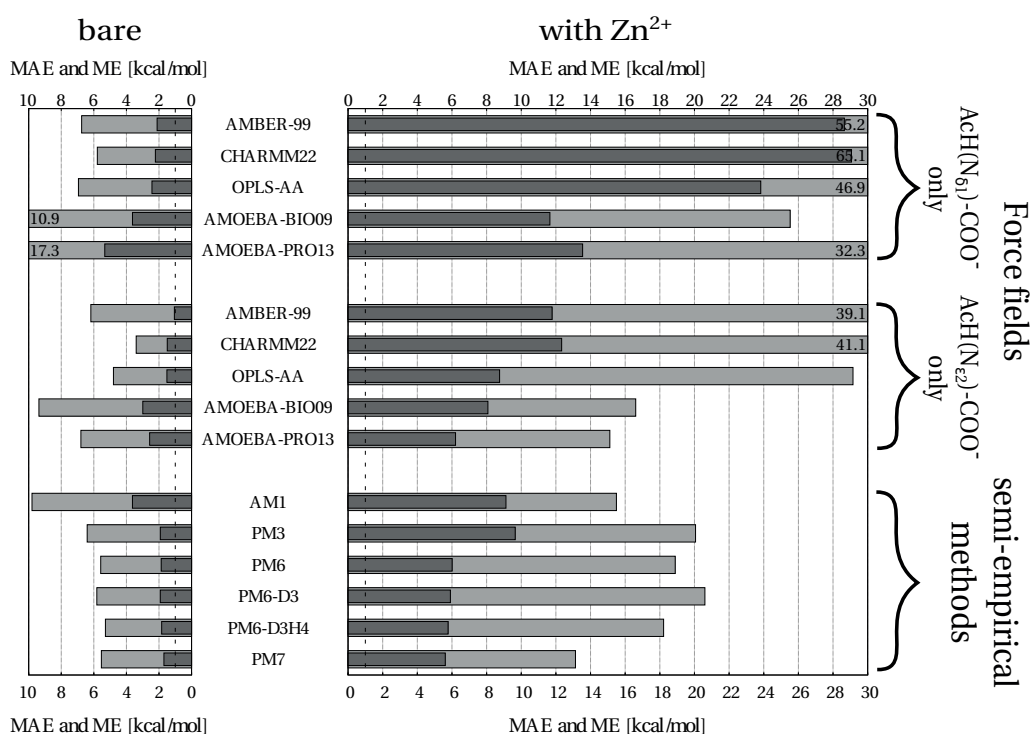
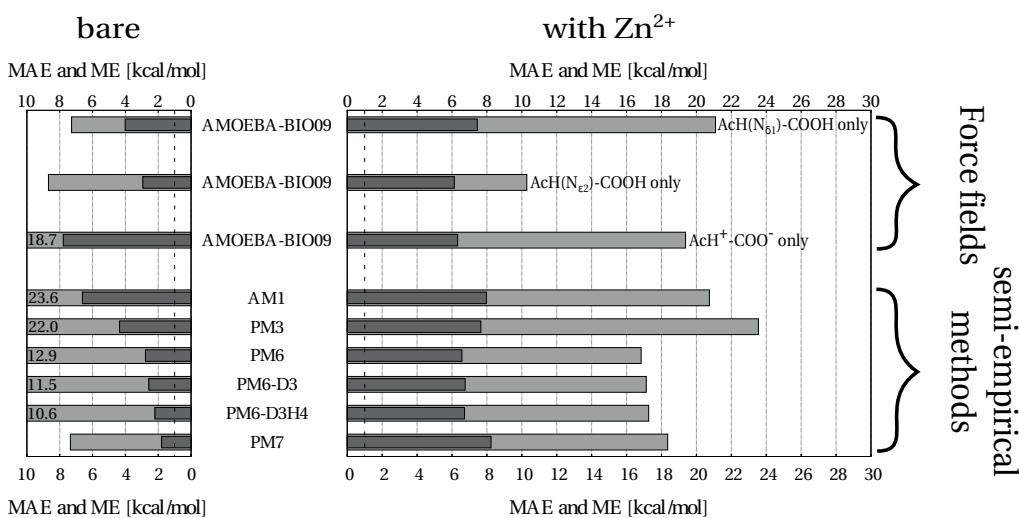
a) AcH-COO<sup>-</sup> (neutral imidazole)b) AcH-COOH (neutral imidazole) / AcH<sup>+</sup>-COO<sup>-</sup>

Figure 4.6 – MAEs (dark-gray) and MEs (light-gray) following Equations (4.2) and (4.3) for different force fields and semi-empirical methods with respect to DLPNO-CCSD(T) for which counterpoise correction has been done following Equation (4.1) and extrapolation to the complete basis set limit has been done following Equations (2.156) and (2.157). The tackled systems are (a) negatively charged AcH with and without a Zn<sup>2+</sup> cation, and (b) neutral AcH with and without a Zn<sup>2+</sup>. Concerning FFs, the different protonation states have to be treated separately.

With a single  $\text{Zn}^{2+}$  cation present, both FFs and semi-empirical methods show very poor performances. Out of the conventional FFs, OPLS-AA shows the best performance with a still very large MAE of 23.8 kcal/mol for  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$  and 8.7 kcal/mol for  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$ . Polarizable atomic multipole-based FFs perform slightly better with a MAE of 11.7 kcal/mol using AMOEBA-BIO09 for  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$  and a MAE of 6.2 kcal/mol using AMOEBA-PRO13 for  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$ . Semi-empirical methods show a further improvement, with PM7 yielding a MAE of 5.6 kcal/mol and a ME of 13.1 kcal/mol.

As AMOEBA-BIO09 is the only FF available providing parameters out-of-the-box for the neutral carboxyl group ( $-\text{COOH}$ ), FF calculations for systems containing neutral AcH are only laid out using this particular FF, as seen in Figure 4.6(b). Protonation states with a neutral imidazole ring yield a MAE of 4.0 kcal/mol for  $\text{AcH}(\text{N}_{\delta_1})-\text{COOH}$  and 2.9 kcal/mol for  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH}$ . For  $\text{AcH}^+-\text{COO}^-$ , performance is again very poor yielding a MAE of 7.8 kcal/mol and a ME of 18.4 kcal/mol. With a single  $\text{Zn}^{2+}$  cation present, the MAE for AMOEBA-BIO09 is larger than 6 kcal/mol for all three protonation states. Out of the semi-empirical methods, PM6 performs best with a MAE of 6.6 kcal/mol.

### 4.3.5 Benchmarking Standard DFAs and Methods Beyond

Similarly to the previous section, the benchmarking process is re-done for different kinds of DFAs as well as the wavefunction-based MP2 method. Figure 4.7 shows obtained MAEs and MEs calculated according to Equations (4.2) and (4.3) for all systems tackled in this work. Considering bare neutral AcH, see Figure 4.7(a), it is interesting to note that all tested methods already provide a very good accuracy as the MAE is less than 1 kcal/mol in all cases. Out of the applied GGA xc functionals, BLYP+D3 shows best performance with a MAE of 0.4 kcal/mol and a ME of 1.1 kcal/mol. It is interesting to see that the applied long-range dispersion schemes all show significant improvement over the methods excluding such treatment already for systems of such a small size, compare *e.g.* the ME of 3.1 kcal/mol for BLYP with the obtained ME of 1.1 kcal/mol for BLYP+D3. All three different van der Waals treatment methods show a similar performance as the respective obtained MAEs differ by less than 0.1 kcal/mol. Out of the meta-GGA xc functionals, SCAN performs best with a MAE of 0.3 kcal/mol and a ME of 1.0 kcal/mol. Performance of the composite method PBEh-3c is comparable to the bare hybrid xc functional PBE0 with a MAE of 0.8 kcal/mol and a ME of 2.2 kcal/mol. Again, long-range dispersion treatments applied *a posteriori* to the hybrid xc functional calculations improve the performance significantly, compare *e.g.* the ME of 2.7 kcal/mol for B3LYP with the obtained ME of 0.8 kcal/mol for B3LYP+D3. The double hybrid xc functional B3LYP+XYG3 and the wavefunction-based MP2 method perform equally well with a ME of 0.8 kcal/mol and 0.9 kcal/mol, respectively.

With a single  $\text{Zn}^{2+}$  cation present, GGA xc functionals are no longer able to describe the energies within “chemical accuracy”. Best performance is found for PBE+MBD with a MAE of

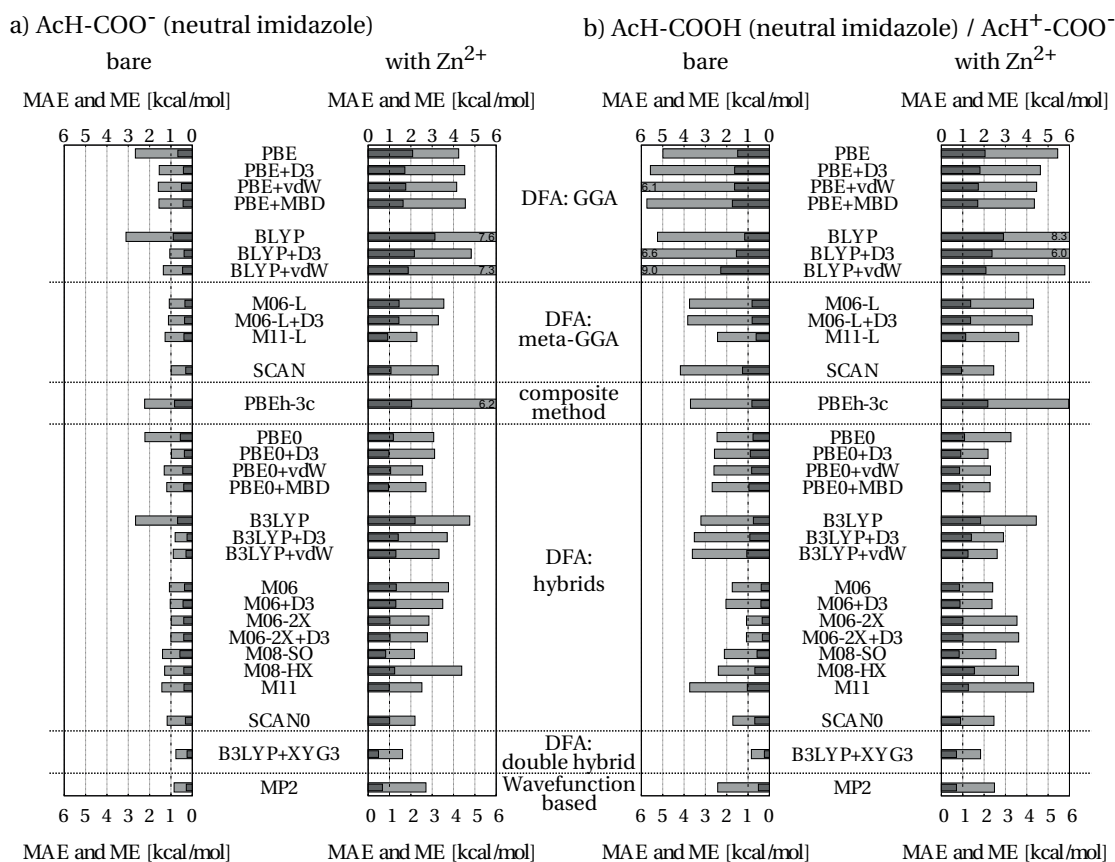


Figure 4.7 – MAEs (dark-gray) and MEs (light-gray) following Equations (4.2) and (4.3) for different standard DFAs, the composite method PBEh-3c, double hybrid DFA B3LYP+XYG3, and the wavefunction-based MP2 method with respect to DLPNO-CCSD(T) for which counterpoise correction has been done following Equation (4.1) and extrapolation to the complete basis set limit has been done following Equations (2.156) and (2.157). The tackled systems are (a) negatively charged AcH with and without a Zn<sup>2+</sup> cation, and (b) neutral AcH with and without a Zn<sup>2+</sup>.

1.6kcal/mol and a ME of 4.6kcal/mol. Meta-GGA xc functionals already yield a big improvement as the M11-L xc functional yields a MAE of 0.9kcal/mol. The composite method PBEh-3c is not sufficient to describe energies of such systems accurately enough as the MAE is found to be 2.0kcal/mol and a rather large ME of 6.2kcal/mol is obtained. Hybrid xc functionals provide a generally more accurate energetic description as PBE0+D3, PBE0+MBD, M06-2X, M06-2X+D3, M08-SO, M11, and SCAN0 yield MAEs within 1.0kcal/mol. The wavefunction-based MP2 method yields a MAE of 0.7kcal/mol. Out of all methods, the double hybrid xc functional B3LYP+XYG3 performs best with a MAE of 0.5kcal/mol and a ME of 1.6kcal/mol.

For neutral AcH, see Figure 4.7(b), the benchmarking process is much more challenging as three different protonation states are considered, as well as minima that differ in energy by up to more than 50kcal/mol. Hence, GGA xc functionals are not able to yield MAEs within

“chemical accuracy”. Best performance is seen for BLYP with a MAE of 1.2 kcal/mol and a ME of 5.2 kcal/mol. Meta-GGA xc functionals already show a big improvement with M11-L giving the best performance with a MAE of 0.6 kcal/mol. The composite method PBEh-3c also yields a small MAE of 0.8 kcal/mol while the corresponding ME of 3.7 kcal/mol indicates that larger energetic deviations are possible for individual conformers. Hybrid xc functionals again perform very well as all MAEs are within 1.0 kcal/mol. Best performance is found for M06-2X with a MAE of 0.3 kcal/mol and a ME of 1.1 kcal/mol. Out of all methods, best performance is again found for B3LYP+XYG3 with a MAE of 0.2 kcal/mol and a ME of 0.8 kcal/mol.

With a single  $\text{Zn}^{2+}$  cation present, performance of the methods is comparable to  $\text{AcH}^+ - \text{COO}^- + \text{Zn}^{2+}$ . GGA xc functionals all yield a MAE above 1 kcal/mol. In order to reach “chemical accuracy” one needs to rely on meta-GGA where the SCAN functional yields a MAE of 0.9 kcal/mol. Out of the hybrid xc functionals, PBE0+D3, PBE0+vdW<sup>TS</sup>, PBE+MBD, M06, M06+D3, M06-2X, M06-2X+D3, M06-SO, and SCAN0 yield MAEs within 1.0 kcal/mol. Out of all methods, best performance is again found for B3LYP+XYG3 with a MAE of 0.7 kcal/mol and a ME of 1.8 kcal/mol.

### 4.3.6 Considering Calculation Times

For applications, one not only needs to consider the accuracy of a particular method, but also the required computational costs and times. All FF and semi-empirical calculations in this work have been laid out on a single CPU core and took between 0.1 s and 0.3 s per single-point energy evaluation. Timings of these methods are all similar due to the small size of the benchmark systems. Because of the fast timings of energy evaluations, conventional FFs are applied if an excessive amount of single-point energy evaluations is required, *e.g.* for molecular dynamics simulations or conformational searches. However, problematic is the lack of well-tested parameterizations for special cases like cations, as seen in this work where this energy description model was found only acceptable for bare neutral AcH. One should therefore generally cross-check with other more accurate methods.

Concerning DFT calculations, timings depend on the applied xc functional, used basis sets, the system, applied convergence criteria and the implementation of the method itself. On a machine with 32 CPU cores and for the system of  $\text{AcH}(\text{N}_{\delta 1}) - \text{COOH} + \text{Zn}^{2+}$ , it took 40 s on average for a single-point energy calculation including force evaluations with FHI-aims applying the GGA xc functional PBE, using `tier 2` basis sets and `really_tight` settings. Using the SCAN xc functional and the M11-L meta-GGA xc functional with the same settings took 77 s (without force evaluations) and 107 s (including force evaluations) on average, respectively. Calculations for the two best performing hybrid xc functionals M08-SO and SCAN0 took 848 s (including force evaluations) and 602 s (without force evaluations) on average using the same settings.

However, for most DFT production purposes one would not rely on computationally costly, yet very accurate, `really_tight` settings, as done in this work. For standard cases, `tight`

settings in combination with `tier 2` basis sets already provide meV-level accurate energy differences [264], *i.e.* within 0.02 kcal/mol. Indeed, repeating the procedure for the PBE, BLYP, PBE0, and B3LYP xc functionals but using `tight` settings yields virtually identical results. On a machine with 32 CPU cores and for the system of  $\text{AcH}(\text{N}_{\delta_1})-\text{COOH} + \text{Zn}^{2+}$ , computational time gets then reduced from 40s to 24s on average for a single-point energy calculation including force evaluations applying the GGA xc functionals PBE. Similarly for the hybrid xc functional PBE0, average calculation times of 738s with `really_tight` settings get reduced to 726s with `tight` settings.

The composite method PBEh-3c that gave MAEs within “chemical accuracy” for the systems without a  $\text{Zn}^{2+}$  cation, took 213s on average for a single-point energy calculation on a single CPU core using ORCA, which is very moderate in computational costs. The most accurate method across all systems and protonation states, B3LYP+XYG3, took 792s on average for a single-point energy evaluation on a machine with 32 CPU cores using FHI-aims. While MAEs for MP2 are comparable with B3LYP+XYG3 and within “chemical accuracy”, energy evaluation times are much larger due to the large basis sets required for accurate predictions. On a machine with 32 CPU cores, it took 2276s on average for an MP2 energy calculation using ORCA and the ZORA-def2-QZVPP basis set.

## 4.4 Conclusions

The goodness of commonly applied levels of theory, *i.e.* force fields, semi-empirical methods, density-functional approximations (DFAs), and wavefunction-based methods were examined with respect to high-level coupled-cluster calculations. To that end, benchmark systems consisting of either a bare acetylhistidine or microsolvated with a  $\text{Zn}^{2+}$  cation were (i) conformationally sampled by performing a global energy minimum search combining both FF and DFA, and (ii) obtained conformational minima were used for benchmarking against DLPNO-CCSD(T) single-point energy-calculations.

For bare negatively charged AcH, the obtained energy hierarchies on the hybrid DFA level showed that the protonation state of  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^-$  is energetically favorable compared to  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^-$  as the respective global minima differ by 14.3 kcal/mol in energy. The situation is reversed with a single  $\text{Zn}^{2+}$  cation present: the protonation state of  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$  is energetically preferred as the respective global minima differ by 18.2 kcal/mol. Considering bare neutral AcH, the two protonation states of  $\text{AcH}(\text{N}_{\delta_1})-\text{COOH}$  and  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COOH}$  yield global minima that are similar in energy. With a single  $\text{Zn}^{2+}$  cation present,  $\text{AcH}(\text{N}_{\epsilon_2})-\text{COO}^- + \text{Zn}^{2+}$  is energetically preferred to the other two protonation states of  $\text{AcH}^+ - \text{COO}^- + \text{Zn}^{2+}$  and  $\text{AcH}(\text{N}_{\delta_1})-\text{COO}^- + \text{Zn}^{2+}$ , as the global minima differ by 16.9 kcal/mol and 29.5 kcal/mol in energy, respectively.

The benchmarking process, based on single-point energy calculations and assessed by means of MAEs and MEs, revealed that force fields and semi-empirical methods are generally not reliable enough for an energetic description of these systems within “chemical accuracy” of

## **Chapter 4. Energetics and Benchmark of Energy Methods of Acetyl-Histidine with Zn<sup>2+</sup>**

---

1 kcal/mol. While GGA xc functionals like PBE and BLYP, as well as the composite method PBEh-3c have problems in their energetic description for systems containing a Zn<sup>2+</sup> cation, it is possible to reach “chemical accuracy” for all systems already using the meta-GGA SCAN xc functional. Hybrid xc functionals perform generally well with MAEs within 1 kcal/mol for most of them. Out of the hybrid xc functionals, best performance is shown for M06-SO and SCAN0. Out of all tested methods, the double hybrid xc functional B3LYP+XYG3 resembles the benchmark method DLPNO-CCSD(T) best with a MAE of 0.7 kcal/mol and a ME of 1.8 kcal/mol. While MP2 performs similarly as B3LYP+XYG3, computational costs, *i.e.* timings, are increased by a factor of 4 in comparison due to the large basis sets required for accurate results.

## **5 Force Field Parameterization Using Regularized Linear Regression**

### 5.1 Motivation

In Chapter 4, the energetics of cation-peptide interactions were investigated and the goodness of commonly applied theoretical levels of theory were assessed. One major finding was described in detail in Section 4.3.4: Conventional FFs showed generally very poor performances therein. This is of particular interest since computational costs of such empirical potentials are low, as *e.g.* described in Subsection 4.3.6, rendering them in principle desirable for tasks like simulating large realistic biomolecular systems or for large-scale structure searches that require enormous amounts of single-point energy evaluations. However, the findings described in Chapter 4 severely limit the applicability of FFs for reliable quantitative predictions. The question furthermore remains if the discrepancy in the energetic description mainly stems from the usage of standard FF parameters that have originally been derived using different systems, or is it the FF formulation of the potential energy term itself that limits an accurate energetic representation.

Many efforts have been made for deriving general-purpose FF parameters for describing systems including metal cations. As an example, parameters of the Lennard-Jones 12-6 potential (see Equation 2.4 in Section 2.2 or Equation 2.13 in Subsection 2.3.1) for alkali and alkaline-earth metals have already been derived by Åqvist in 1990 using experimental hydration free energy values [364]. Similarly, Stote and Karplus derived parameters of the Lennard-Jones 12-6 potential for  $\text{Zn}^{2+}$  in 1995 [365]. A variety of different approaches for parameterization for Lennard-Jones and electrostatic interactions with metal ions followed, as for example summarized in detail in Reference [366]. Almost all these attempts have in common that the parameterization and testing is done on systems in solution using either explicit or implicit solvation models, while the focus in this work lies on gas-phase calculations, as laid out in Section 2.2. Furthermore, the classical modeling of metal ions using bonded models has been researched already since the 1960s [366], an approach which commonly requires an *a priori* definition of bonds involving the metal cation in question. Obviously, such approaches severely restrict global conformational search methods like the one described in Section 3.3, independent of their energetic accuracy. In any case, a conventional parameterization approach is commonly a tedious and time-consuming process [367] and hence generally not feasible to be undergone by the end-users themselves.

Taking these points and the findings in Chapters 3 and 4 into account, the idea to be able to adjust parameters of a particular FF for a specific system in question, *e.g.* a certain peptide-cation system in the gas phase, becomes appealing. The minimum initial demand shall thereby be to be able to modify the FF parameters in such a way that the energy hierarchies obtained using DFT (or any other high-level method) are to be reproduced within a certain threshold, *e.g.* within “chemical accuracy” of 1 kcal/mol using MAEs introduced in Subsection 4.2.3. One further important aspect lies on the approach to be rather simplistic in order to be able to be undergone by the end-users themselves. A framework for a machine learning approach that aims to fulfill the described task is to be presented in detail in this chapter. In essence, torsional parameters and (if desired) van der Waals parameters in the potential-energy function  $E_{\text{pot}}$  of



a particular FF, here OPLS-AA, are adjusted by simply fitting  $E_{\text{pot}}$  against high-level energies, *e.g.* from DFT calculations, using different regression methods for a rather small subset out of a large pool of conformers. Because FF parameters are obtained from regression methods using only the potential energy obtained from DFT for a specific system in question, the set-up allows for immediate verification of how well the FF formulation itself is able to describe the potential energy, a venture to be undertaken quantitatively for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup>.

The idea to use a machine-learning approach in order to derive FF parameters “more appropriate” for a specific system in question has been intended in the past. Huang and Roux set up a general method for small molecules that aims to automatically generate parameters of a FF which potential energy function  $E_{\text{pot}}$  is similar to the ones of the AMBER and CHARMM FFs (see Subsection 2.3.1) [368]. A “black box” web server thereof is available [369]. Initial guesses of the parameters are taken from previously developed FFs, and a variety of *ab initio* quantum mechanical calculations like AM1, HF, MP2, and hybrid DFAs is used as target data when optimizing different objective functions in order to generate optimized FF parameters. Dihedral parameters are optimized using 1-dimensional dihedral scans and energies of conformers calculated from quantum mechanical methods. Li, Roux, and coworkers applied a machine learning technique based on a genetic algorithm in order to predict force field parameters using *ab initio* data from quantum mechanics calculations [370]. The concept showed promising results when applied for methanol clusters. Fracchia, Barone, and coworkers developed a statistical procedure that aims to optimize parameters of non-bonded FFs of metal ions in soft matter [371]. Basically, the optimization process is laid out by minimizing the deviations from *ab initio* forces and energies by applying Ridge regression [15–17] and cross-validation techniques. Instead of using a fixed classical FF term (*e.g.* the Lennard-Jones 12-6 potential), a variety of possible models are tested in a systematic comparative study, thus effectively suggesting an “optimized form” of the potential energy function for a particular system in question. For the test case of cations in water, results are promising. Finally, many more machine learning approaches for describing energetics of molecular systems beyond the force field description exist, *e.g.* a machine learning model to predict atomization energies of organic molecules [372], non-linear dimensionality reduction techniques to classify molecular structures and map conformational free energies [373–375], as well as neural networks and Gaussian approximation potentials to represent multidimensional potential energy surfaces [376–381], to name but a few.

After laying out in detail the theoretical background as well as the framework and concept of the approach in Section 5.2, a *proof of principle* is intended in Section 5.3 using the toy model of AcAla<sub>2</sub>NMe, and the more challenging peptide-cation system of AcAla<sub>2</sub>NMe + Na<sup>+</sup> is tackled.

## 5.2 Computational Details and Framework

### 5.2.1 Functional Form and Parameters of Empirical Force Fields

The description of a conventional empirical FF has already been provided in detail for the example of OPLS-AA [12–14] in Subsection 2.3.1. Thus, it is only briefly summarized here in order to highlight the different classes and types of FF parameters that will either be optimized using different regression models (see Subsection 5.2.2) or for which functions thereof will serve as descriptors of the model, as laid out in Subsection 5.2.3. As explained in Subsection 2.3.1, the potential energy  $E_{\text{pot}}^{\text{FF}}(\vec{R}^N)$  of a conventional empirical FF is given as a function of positions  $\vec{R}_1, \dots, \vec{R}_N$  of the  $N$  nuclei of the system. It is commonly written as a sum of energy terms, each of them corresponding to qualitatively different interactions:

$$E_{\text{pot}}^{\text{FF}}(\vec{R}^N) = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{tors}} + E_{\text{vdW}} + E_{\text{Coulomb}}, \quad (5.1)$$

where

$$E_{\text{bonded}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{tors}} \quad (5.2)$$

denotes the “bonded” contributions while

$$E_{\text{non-bonded}} = E_{\text{vdW}} + E_{\text{Coulomb}} \quad (5.3)$$

represents the “non-bonded” contributions. The “bonded” terms are of the following form:

$$E_{\text{bonds}} = \sum_{i < j}^{1-2 \text{ atoms}} K_{ij}^r (r_{ij} - r_{ij}^0)^2, \quad (5.4)$$

$$E_{\text{angles}} = \sum_{i < j}^{1-3 \text{ atoms}} K_{ij}^\theta (\theta_{ij} - \theta_{ij}^0)^2, \quad (5.5)$$

$$E_{\text{tors}} = \sum_{i < j}^{1-4 \text{ atoms}} \left\{ \frac{V_1^{ij}}{2} (1 + \cos(\phi^{ij})) + \frac{V_2^{ij}}{2} (1 - \cos(2\phi^{ij})) + \frac{V_3^{ij}}{2} (1 + \cos(3\phi^{ij})) \right\}. \quad (5.6)$$

The sum in Equation (5.4) is over all 1-2 atoms, *i.e.* pairs of atoms bonded to each other, while the sum in Equation (5.5) is over all 1-3 atoms or bond angles, *i.e.* atoms  $i$  and  $j$  that are separated by two bonds. The potential energy of the bonds and angles is approximated as a harmonic oscillator, *i.e.* as a quadratic function of the displacement of the bond length  $r_{ij}$  from its reference length  $r_{ij}^0$ , or similarly the bond angle  $\theta_{ij}$  from its reference bond angle  $\theta_{ij}^0$ .  $K_{ij}^r$ ,  $r_{ij}^0$ ,  $K_{ij}^\theta$ , and  $\theta_{ij}^0$  are empirical parameters that depend on the atom classes of the participating pairs or triplets of atoms in question. It should be pointed out that the quadratic form of those terms primarily serves to provide a “basic rigid” structural form, *i.e.* ensuring that bonded atoms are always separated with a bond length  $r_{ij}$  near its reference length  $r_{ij}^0$ , similarly for the bond angle  $\theta_{ij}$  and its reference bond angle  $\theta_{ij}^0$ . The flexibility of a functional form that is able to accurately describe energetic properties of different conformers of the same peptide is mainly aimed to be provided by Equation (5.6), *i.e.* the “torsional” term

of the potential energy  $E_{\text{tors}}$ . The sum is thereby over all 1-4 atoms or torsional angles, *i.e.* atoms  $i$  and  $j$  that are separated by three bonds, and the empirical parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  depend on the atom classes of the four atoms defining the torsional angle. Note that all torsional parameters contribute linearly using this functional form, in contrast to *e.g.* the — albeit equivalent — functional form used in the description of the AMBER-99 and CHARMM22 FFs, see Equation (2.12).

Non-covalent, *i.e.* “non-bonded”, inter-atomic Coulomb and vdW interactions are described by Equations (2.14) and (2.13), respectively, *i.e.*

$$E_{\text{Coulomb}} = \sum_{i < j} \frac{q_i q_j}{r_{ij}} f_{ij}, \quad (5.7)$$

$$E_{\text{vdW}} = \sum_{i < j} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] f_{ij}, \quad (5.8)$$

where the sum runs over all pairs of atoms  $i$  and  $j$ . The empirical parameters  $q_i$ ,  $\epsilon_{ij}$ , and  $\sigma_{ij}$  depend on the atom types of the participating atoms or atom pairs in question. As already described in Subsection 2.3.1, the 1-2 and 1-3 interactions are considered to be already implicitly included in their respective “bonded” contributions, and it was found to be necessary to scale the corresponding 1-4 interactions by a factor of  $\frac{1}{2}$  [13]. Hence, the scaling factor  $f_{ij}$  is written as

$$f_{ij} = \begin{cases} 0, & \text{for 1-2 and 1-3 atoms,} \\ \frac{1}{2}, & \text{for 1-4 atoms,} \\ 1, & \text{otherwise.} \end{cases} \quad (5.9)$$

Since the work here focuses on peptide-cation systems in the gas phase, the “non-bonded” terms commonly contribute significantly to the total potential energy, as the interactions between atoms are “undamped” due to the lack of shielding due to solvation. Hence when adjusting or optimizing parameters, it is a natural choice to first focus on those empirical parameters included in these terms, namely the  $\epsilon_{ij}$  and  $\sigma_{ij}$  of the vdW term as well as the products of atomic partial charges  $q_i q_j$  in the Coulomb term. Furthermore, this might also impact the covalent structure of the peptide, *e.g.* previous studies have shown that peptide-cation interactions may enforce non-standard torsions [335, 382]. Finally, with additionally adding an optimization process for the torsional parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$ , the goal is to be able to yield an accurate yet computationally cheap energetic FF description for a particular peptide-cation system in question. The concept for that is explained in Subsection 5.2.3 and depends on the use of regularized linear regression models, hence they are laid out in the following chapter.

### 5.2.2 Regularized Linear Regression: Ridge Regression and LASSO

The work within this chapter relies on regularized linear regression models that in themselves depend on the common multiple linear regression model. Assuming a data set of sample size  $n$  that consists of collected matched pairs  $(\mathbf{x}_i, Y_i)$ ,  $i = 1, \dots, n$ , where  $\mathbf{x}_i$  denotes the input vector, *i.e.* the vector  $\mathbf{x}_i$  containing  $p$  predictor variables (sometimes also named input or feature variables)  $x_{il}$ ,  $l = 1, \dots, p$ , such that

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}, \quad (5.10)$$

and  $Y_i$  is the  $i$ -th output value or response variable, the multiple linear model then specifies a linear relationship between  $\mathbf{x}_i$  and the expected value  $E[Y_i]$  that in turn equals the mean response  $\mu(\mathbf{x}_i)$ , *i.e.*  $E[Y_i] = \mu(\mathbf{x}_i) = \mu(x_{i0} = 1, x_{i1}, x_{i2}, \dots, x_{ip})$ . In other words, the mean response  $\mu(\mathbf{x}_i) = \mu_i$  is modeled as a linear predictor such that

$$\mu_i = \mu(\mathbf{x}_i) = E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad i = 1, \dots, n, \quad (5.11)$$

where the  $\beta_k$ ,  $k = 0, \dots, p$ , denote the  $p + 1$  linear regression coefficients. Making use of matrix notation by writing the response vector  $\mathbf{Y}$  as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad (5.12)$$

the vector  $\boldsymbol{\beta}$  containing the regression coefficients as

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad (5.13)$$

and the design matrix  $\mathbf{X}$  as

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad (5.14)$$

Equation (5.11) then translates to

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}, \quad (5.15)$$

with the individual entries of  $E[\mathbf{Y}]$  corresponding to

$$E[Y_i] = \mu_i = \mu(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i = 1, \dots, n. \quad (5.16)$$

The regression coefficients are commonly estimated by employing the method of least squares [383, 384], that is constructing and minimizing the objective function  $\mathcal{D}$ , given by

$$\begin{aligned} \mathcal{D} &= \sum_{i=1}^n [Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}]^2 \\ &= \sum_{i=1}^n \left[ Y_i - \left( \beta_0 + \sum_{l=1}^p \beta_l x_{il} \right) \right]^2, \end{aligned} \quad (5.17)$$

with respect to the regression coefficients  $\beta_k, k = 0, \dots, p$ . This produces a system of normal equations [385]

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y} \quad (5.18)$$

whose solution yields the unique least squares estimators  $\hat{\boldsymbol{\beta}}$  [386]. Assuming the rank of the design matrix  $\mathbf{X}$  equal to  $p + 1$  such that  $(\mathbf{X}^\top \mathbf{X})^{-1}$  is well defined, the solution for the least squares estimators  $\hat{\boldsymbol{\beta}}$ , *i.e.* the solution of Equation (5.18), is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}. \quad (5.19)$$

The vector  $\hat{\mathbf{Y}}$  containing the fitted values  $\hat{Y}_i, i = 1, \dots, n$ , is thus given by

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}. \end{aligned} \quad (5.20)$$

Finally, the residual sum of squares  $RSS$  is written as [385]

$$\begin{aligned} RSS &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \mathbf{Y}^\top (\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{Y}, \end{aligned} \quad (5.21)$$

where  $\mathbf{I}$  denotes the identity matrix.

In case of high correlations among predictor variables or just a large number of predictors, say for  $p \gtrsim n$ , the multiple linear regression model may be ill-posed, meaning the matrix  $\mathbf{X}^\top \mathbf{X}$  in Equation (5.19) may appear almost singular, often labeled *ill-conditioned*, hence resulting in a numerically unstable inverse matrix  $(\mathbf{X}^\top \mathbf{X})^{-1}$  that in turn yields numerically unstable least squares estimates  $\hat{\boldsymbol{\beta}}$  [387]. One possibility to account for that is to make use of regularization [388] or shrinkage regression [389]: By penalizing the regression parameters, *i.e.* by artificially shrinking them towards the origin, one aims to stabilize them. Formally, this

## Chapter 5. Force Field Parameterization Using Regularized Linear Regression

may be achieved by adding a penalty term to the objective function  $\mathcal{D}$  in Equation (5.17), *i.e.*

$$\mathcal{D} = \sum_{i=1}^n \left[ Y_i - \left( \beta_0 + \sum_{l=1}^p \beta_l x_{il} \right) \right]^2 + \mathcal{F}(\beta_0, \dots, p), \quad (5.22)$$

where  $\mathcal{F}(\beta_0, \dots, p)$  denotes the Tikhonov factor [390], a positive penalty function on the regression coefficients  $\beta_k, k = 0, \dots, p$ . A common choice for  $\mathcal{F}$  employs a quadratic form and results in the Ridge regression model [15–17], *i.e.*

$$\mathcal{D} = \sum_{i=1}^n \left[ Y_i - \left( \beta_0 + \sum_{l=1}^p \beta_l x_{il} \right) \right]^2 + \lambda \sum_{l=1}^p \beta_l^2, \quad (5.23)$$

where  $\lambda$  denotes a regularization or tuning parameter, and  $\lambda \geq 0$ . The tuning parameter  $\lambda$  thereby acts as a Lagrange multiplier within this constrained optimization problem [391]. Thus, equivalently to Equation (5.23), one may re-formulate the problem [392]:

$$\text{minimize } \sum_{i=1}^n \left[ Y_i - \left( \beta_0 + \sum_{l=1}^p \beta_l x_{il} \right) \right]^2 \text{ subject to } \sum_{l=1}^p \beta_l^2 \leq \tilde{\lambda}. \quad (5.24)$$

For the sake of simplicity, one may center and scale the predictor variables  $x_{ik}, i = 1, \dots, n, k = 0, \dots, p$ , yielding the so-called z-scores  $z_{ik}$  given by [385]

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}, \text{ with } \bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}, \text{ and } s_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}, \quad (5.25)$$

with  $\bar{x}_k$  just denoting the arithmetic mean of the  $k$ -th predictor, with  $s_k$  denoting the corresponding standard deviation. Using this notation, the estimator of  $\beta_0$  is just the arithmetic mean of the output variables  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ . Using the centered response variables  $U_i = Y_i - \bar{Y}$ , the corresponding centered response vector  $\mathbf{U}$  given by

$$\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}, \quad (5.26)$$

the z-scores design matrix  $\mathbf{Z}$  given by

$$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix}, \quad (5.27)$$

and the vector  $\boldsymbol{\beta}$  again containing the regression coefficients, *i.e.*

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad (5.28)$$

the objective function  $\mathcal{D}$  in Equation (5.23) is simply written in matrix notation as

$$\mathcal{D} = (\mathbf{U} - \mathbf{Z}\boldsymbol{\beta})^\top (\mathbf{U} - \mathbf{Z}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}. \quad (5.29)$$

Minimizing  $\mathcal{D}$  yields the normal equations given by [385]

$$(\mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I})\boldsymbol{\beta} = \mathbf{Z}^\top \mathbf{U}, \quad (5.30)$$

whose solution yields the unique estimators  $\hat{\boldsymbol{\beta}}$ . Comparing with Equation (5.18) for the case of conventional multiple linear regression, the form is the same except a diagonal “ridge” ( $\lambda \mathbf{I}$ ) has been added to the  $\mathbf{Z}^\top \mathbf{Z}$  matrix, essentially stabilizing it so effectively to always ensure the existence of an inverse. Hence, the solution for the estimators  $\hat{\boldsymbol{\beta}}$  always exists:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I})^{-1} \mathbf{Z}^\top \mathbf{U}. \quad (5.31)$$

Finally, the vector  $\hat{\mathbf{U}}$  containing the Ridge predicted values  $\hat{U}_i, i = 1, \dots, n$ , is given by

$$\begin{aligned} \hat{\mathbf{U}} &= \mathbf{Z}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I})^{-1} \mathbf{Z}^\top \mathbf{U}. \end{aligned} \quad (5.32)$$

Obviously, the Ridge estimator is biased. For  $\lambda \rightarrow 0$ , the bias vanishes. For  $\lambda \rightarrow \infty$ , the bias is so large that the estimates shrink to zero. Hoerl and Kennard [17] showed that there always exists some  $\lambda (> 0)$  such that the residual sum of squares *RSS* (see Equation (5.21)) is smaller than for the ordinary least squares estimator. However,  $\lambda$  is commonly unknown *a priori* because it depends on the regression coefficients  $\boldsymbol{\beta}$  themselves. In practice, this usually means that one needs to conduct  $\lambda$  in an exploratory manner. For example, one may simply “train”  $\lambda$  using a set of training data, *i.e.* essentially minimizing the *RSS* with respect to  $\lambda$  without too much shrinking the regression coefficients  $\boldsymbol{\beta}$  towards the origin. For selecting  $\lambda$  from a single set of data, different approaches exist, *e.g.* a standalone estimate [393], a *ridge trace* [17], or a form of *cross-validation* [394].

Instead of the “ $L_2$ ” penalty term in Equation (5.23), one may instead use a “ $L_1$ ” penalty term in the regularization objective function  $\mathcal{D}$ , resulting in the LASSO (least absolute shrinkage and selection operator) regression model [18], *i.e.*

$$\mathcal{D} = \sum_{i=1}^n \left[ Y_i - \left( \beta_0 + \sum_{l=1}^p \beta_l x_{il} \right) \right]^2 + \lambda \sum_{l=1}^p |\beta_l|. \quad (5.33)$$

Similarly to Equation (5.24), one may re-formulate the problem to

$$\text{minimize } \sum_{i=1}^n \left[ Y_i - \left( \beta_0 + \sum_{l=1}^p \beta_l x_{il} \right) \right]^2 \text{ subject to } \sum_{l=1}^p |\beta_l| \leq \tilde{\lambda}. \quad (5.34)$$

In contrast to Ridge regression, there exists no closed-form expression for the estimators  $\hat{\beta}$ . Instead, a numerical solution is required. The applied form of shrinkage generally yields so-called *sparse* solutions, meaning some estimators  $\hat{\beta}_l, l = 1, \dots, p$  are explicitly set to zero, meaning that in general only a subset of the original estimators  $\hat{\beta}_l$  is yielded from the “ $L_1$ ”penalized fit. In other words, the LASSO regularization method not only serves as a form of shrinkage regression, but also as a *de facto* variable selector, as it completely suppresses the low-impact predictor variables [395]. Hence, the LASSO regression model is often applied in case of a large number of predictor variables or overdetermination [385], although Ridge regression usually yields more predictive capability in case of high multicollinearity among predictor variables [18].

### 5.2.3 The “Framework For Adjusting Force Fields Using Regularized Regression” (FFAFFURR)

A machine-learning approach is intended in order to adjust parameters from an already existing FF, here OPLS-AA which its functional form has been described in Subsection 5.2.1, such that the energy hierarchies obtained using DFT (or any other high-level method) are to be reproduced within a certain threshold, *e.g.* within “chemical accuracy” of 1 kcal/mol using MAEs introduced in Subsection 4.2.3. This minimum initial demand is just thought to be a “first stepping stone” when keeping in mind that at some point in the future the approach should be extended to work not only with conformational energies but also including forces. In fact, the same argument applies when justifying the usage of the rather rigid functional form of a conventional FF in the first place, instead of any other functional form probably more suitable to accurately describe conformational energy hierarchies. Although the (in part) physical motivation behind the rather simple scheme of conventional FFs might appear appealing, it is anything but clear if the FF formulation itself is capable of accurately describing the energetics of conformers. Instead, inaccuracies in the energetic description of a conventional FF for a particular system are often alluded to insufficient parameterization [396]. An automatized parameterization approach like the one presented here might help to immediately verify how well the FF formulation itself is able to describe the conformational energy of conformers, as FF parameters are obtained from regression methods using only the potential energy obtained from DFT for a specific system in question.

When setting up the underlying framework, three practical points were taken into consideration: (i) The framework should be sufficiently simple for an end-user to set-up. (ii) It should be easy to extend for usage with other FF parameters or functional forms. (iii) It should run without explicitly depending on third-party programs, although it is evident that the FF



parameters and energies calculated at the DFT level must be read in as input. Similarly, the output FF parameters should be provided in a way to immediately be usable by a molecular modeling package. Because the TINKER program was already used extensively in Chapters 3 and 4 of this work, it is a natural choice to focus on this particular software. To be more precise, in this work version 7.1.2 of the TINKER package is used. On the other hand, input from DFT calculations are provided using FHI-aims. To summarize, the “Framework For Adjusting Force Fields Using Regularized Regression” (FFAFFURR)<sup>1</sup>, written in Python, acts as a “wrapper” between the molecular modeling package TINKER and the *ab initio* molecular simulations package FHI-aims in a sense that output files produced by these programs serve as input in order to read in all required informations (initial FF parameters, conformational energies, *etc.*) for adjusting FF parameters that are then provided as output immediately capable of being processed further by TINKER.

In the following three subsections, the framework and its underlying approach will be laid out in detail, in particular input, concept, and output, respectively.

### FFAFFURR: Input

Because the goal is to adjust existing parameters of the OPLS-AA FF, it is reasonable to provide a standardized listing of parameters as input. This is achieved for any system already set-up with TINKER, *e.g.* when using the standard OPLS-AA FF parameters that are distributed with the package, by using the analyze tool, *i.e.* by issuing the following command:

Listing 5.1

```
1 analyze -k sample.key sample.xyz P ALL > ffaffurr.input.originalFF
```

The `sample.key` file thereby denotes the keyword parameter file in TINKER that most importantly contains the location of the potential energy parameter file, *e.g.* `oplsaa.prm` for the standard OPLS-AA FF distributed with the TINKER package. The `sample.xyz` file is the basic TINKER coordinate file type. The standardized listing of parameters is redirected into the `ffaaffurr.input.originalFF` file that needs to be provided as input to FFAFFURR by placing it in the same directory as the Python file `ffaaffurr.py`. In a similar fashion, a standardized connectivity list for each of the atoms may be generated by issuing the following command:

Listing 5.2

```
1 analyze -k sample.key sample.xyz C ALL > ffaffurr.input.interactionsFF
```

As before, the `ffaaffurr.input.interactionsFF` file needs to be provided as input by placing it in the same directory as the Python file `ffaaffurr.py`.

Third, the input file `ffaaffurr.input.FHI-aims-logfiles` contains a list of FHI-aims-specific

<sup>1</sup>The code is available free of charge and can be downloaded from:  
<https://github.com/FHIBioGroup/ffaaffurr-dev>

## Chapter 5. Force Field Parameterization Using Regularized Linear Regression

---

output files produced when calculating single-point DFT energies. Obviously, these files must be produced for a set of conformers that serves as training data.

Finally, the input file `ffaffurr.input` contains the “switches” that control the behavior of the framework, *e.g.* what kind of parameters are to be adjusted or what regression model to use. Details for the different kinds of FF parameters are provided in the following.

### FFAFFURR: Concept of Adjusting the Force Field Parameters

The bonds and angles contributions to the potential energy in the OPLS-AA FF formulation are given in Equations (5.4) and (5.5), respectively. As laid out in Subsection 5.2.1, their quadratic form primarily serves to provide the peptide in question its “basic rigid” structural form, as opposed to accurately describe energetic properties of the molecular system. This is why the “spring” parameters  $K_{ij}^r$  and  $K_{ij}^\theta$  are unaltered in this study, while the focus lies on the torsional and non-bonded parameters. However, the equilibrium parameters  $r_{ij}^0$  and  $\theta_{ij}^0$  between pairs or triplets of atoms may be adjusted by simply averaging over all corresponding atomic pairwise distances of the same pair or triplet of atom classes over all FHI-aims-specific input files. Obviously, this is only useful if the input structures were geometry optimized beforehand, *i.e.* they must be situated in a local minimum on the PES calculated at the DFT level of theory.

The Coulomb contribution to the potential energy is given in Equation (5.7). Within FFAFFURR, it is possible to estimate the atomic partial charge parameters  $q_i$  by assigning them to either Hirshfeld charges or ESP charges calculated with FHI-aims. Hirshfeld atomic charges are thereby derived based on the Hirshfeld partitioning scheme [223–225] introduced in Subsection 2.3.6. The Hirshfeld atomic charge  $q_i$  of atom  $i$  is simply given by

$$q_i = Z_i - \int \rho_i(\vec{r}) d\vec{r}, \quad (5.35)$$

where  $Z_i$  denotes the corresponding atomic number, and  $\rho_i(\vec{r})$  is the associated electron density of atom  $i$  given by

$$\rho_i(\vec{r}) = w_i(\vec{r})\rho(\vec{r}), \quad (5.36)$$

where  $\rho(\vec{r})$  denotes the total electron density and  $w_i(\vec{r})$  is the Hirshfeld atomic partitioning weight defined in Equation (2.130). ESP charges, on the other hand, are commonly derived from *ab initio* or semi-empirical calculations by fitting the partial charges to reproduce the electrostatic potential (ESP) [397–399]. Within FHI-aims, a simple method is implemented [400]: The electrostatic potential  $V_{\text{DFT}}$  is evaluated at a sufficiently high number of grid points within a defined a spatial region, *i.e.* at a particular grid point  $\vec{r}$ , the electrostatic potential  $V_{\text{DFT}}(\vec{r})$  is calculated as

$$V_{\text{DFT}}(\vec{r}) = \sum_i \frac{Z_i}{|\vec{r} - \vec{R}_i|} - \int \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r}', \quad (5.37)$$

where the sum in the first term is over all atoms  $i$  with their corresponding atomic numbers

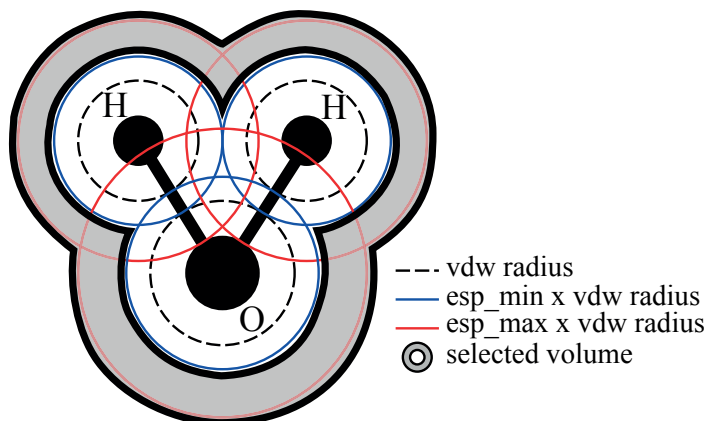


Figure 5.1 – Schematic illustration of the selected volume that confines the grid points at which the electrostatic potential (ESP) is evaluated. Reproduced from Reference [400] with permission from B. Bieniek.

$Z_i$  and positions  $\vec{R}_i$ . The second term is just the Hartree potential from Equation (2.83). The spatial region is thereby defined in terms of multiples of the vdW radii of the atoms, *i.e.* all grid points are situated inside spheres confined by minimal and maximal multiples of the vdW radii. The situation is depicted in Figure 5.1. Values of the vdW radii were taken from previously tabulated data [39, 401]. Within this work, the default minimal and maximal multiples of the vdW radii of 5 and 8 are used throughout. When expressing the electrostatic potential (ESP)  $V_{\text{ESP}}$  in terms of atomic partial charges, *i.e.* the ESP charges  $q_i$ , located at the atomic positions  $\vec{R}_i$  as

$$V_{\text{ESP}}(\vec{r}) = \sum_i \frac{q_i}{|\vec{r} - \vec{R}_i|}, \quad (5.38)$$

one may evaluate the ESP charges  $q_i$  using a simple least-squares fit. The constraint of constant total charge  $q_{\text{tot}} = \sum_i q_i$  is thereby taken into account by applying the method of Lagrange multipliers [115] to minimize the objective function

$$\mathcal{F} = \sum_k^{\text{grid points}} (V_{\text{DFT}}(\vec{r}_k) - V_{\text{ESP}}(\vec{r}_k))^2 - \lambda_q \left( q_{\text{tot}} - \sum_i q_i \right)^2. \quad (5.39)$$

Independently of using the Hirshfeld partitioning scheme or the ESP method, the final atomic partial charges are derived by averaging over all corresponding atoms of the same atom type and over all input structures provided with the FHI-aims-specific input files.

The vdW contribution to the potential energy is given in Equation (5.8). Using FFAFFURR, it is possible to estimate the interatomic pairwise  $\sigma_{ij}$  parameters by using the atomic Hirshfeld partitioning scheme that has already been used in the Tkatchenko-Scheffler vdW<sup>TS</sup> model explained in Subsection 2.3.6. Applying the concept of van der Waals radii, one may write an

equivalent formulation of Equation (5.8), namely

$$E_{\text{vdW}} = \sum_{i < j} \epsilon_{ij} \left[ \left( \frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^6 \right] f_{ij}, \quad (5.40)$$

where the atomic distance  $R_{ij}^{\text{min}}$  at which the vdW potential is at its minimum is just estimated as the sum of the corresponding effective atomic van der Waals radii that in turn are expressed in Equation (2.128). Comparing Equations (5.8) and (5.40) then immediately yields

$$\sigma_{ij} = 2^{-1/6} R_{ij}^{\text{min}}. \quad (5.41)$$

In order to adjust the vdW parameters  $\epsilon_{ij}$  such that the form of the FF potential more accurately describes conformational energy hierarchies obtained from DFT calculations, it is intended to make use of regression models laid out in Subsection 5.2.2. Assuming a training set of conformers of sample size  $n$ , the idea is to use DFT calculations as target data. In particular, the energetic dispersion correction  $E^{\text{vdW,DFT}}$  that is evaluated *a posteriori*, *i.e.* after the self-consistent Kohn-Sham treatment in DFT (see Equation (2.119)), may serve as an appropriate response. As laid out in detail in Subsection 2.3.6, two *a posteriori* vdW schemes are implemented in FHI-aims, namely the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> model and the many-body dispersion scheme MBD. It is possible to adjust the vdW parameters  $\epsilon_{ij}$  using either one of the two schemes in FFAFFURR as target data. In other words, the response vector in Equation (5.12) is just

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{\tilde{i}} \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} E_1^{\text{vdW,DFT}} \\ \vdots \\ E_{\tilde{i}}^{\text{vdW,DFT}} \\ \vdots \\ E_n^{\text{vdW,DFT}} \end{pmatrix}, \quad (5.42)$$

where  $E_{\tilde{i}}^{\text{vdW,DFT}}$  denotes the *a posteriori* evaluated energetic vdW contribution (using either the vdW<sup>TS</sup> or MBD model) of conformer  $\tilde{i}$ . The vector  $\boldsymbol{\beta}$  containing the regression coefficients given in Equation (5.13) is then written as

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} E^{\text{offset}} \\ \epsilon_{11} \\ \epsilon_{12} \\ \vdots \end{pmatrix}, \quad (5.43)$$

*i.e.* the vector  $\boldsymbol{\beta}$  contains all possible  $\epsilon_{ij}$  parameters as regression coefficients attributed to their individual pairs of types of atoms within the description of the OPLS-AA FF, and the “intercept”  $\beta_0$  is just an arbitrary energetic potential offset  $E^{\text{offset}}$ . Taking into account the formulation of the vdW contribution to the potential energy in the OPLS-AA FF description given in Equation (5.8), the vector  $\mathbf{x}_{\tilde{i}}$  containing the predictor variables (see Equation (5.10))

is immediately written as

$$\mathbf{x}_{\tilde{i}} = \begin{pmatrix} 1 \\ x_{\tilde{i}1} \\ x_{\tilde{i}2} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 \\ \left( \sum_{11} 4 \left[ \left( \frac{\sigma_{11}}{r_{11}} \right)^{12} - \left( \frac{\sigma_{11}}{r_{11}} \right)^6 \right] f_{11} \right)_{\tilde{i}} \\ \left( \sum_{12} 4 \left[ \left( \frac{\sigma_{12}}{r_{12}} \right)^{12} - \left( \frac{\sigma_{12}}{r_{12}} \right)^6 \right] f_{12} \right)_{\tilde{i}} \\ \vdots \end{pmatrix}, \quad (5.44)$$

where the sums indicate summations over all pairs of atoms with the same pairs of atom types. Using these definitions, the approach for estimating the vdW parameters  $\epsilon_{ij}$  using different regression models is then straightforward and has already been laid out in Subsection 5.2.2 where the objective functions to minimize for conventional multiple linear regression, Ridge regression, and the LASSO have been given in Equations (5.17), (5.23), and (5.33), respectively. Within the FFAFFURR framework, calculations using the different regression models are laid out by making use of Python's `scikit-learn` [402] library. Obviously, the choice of using *a posteriori* calculated energetic vdW contributions  $E^{\text{vdW,DFT}}$  as target data is kind of arbitrary, although its availability as a separate term to the total energy is an appealing feature. However, as explained in Subsection 2.3.6, the term does depend on the specific xc functional used in the DFT calculation, as the short-range region is mostly described by the underlying DFA. For example, in the  $\text{vdW}^{\text{TS}}$  model this is intended to be accounted for by using the Fermi-type damping function given in Equation (2.127). While the scaling factor  $f_{ij}$  (see Equation (5.9)) within the description of the OPLS-AA FF might in part resemble a similar behavior, one should always be aware of this discrepancy between the different formulations and applied DFAs. Of course, in principle other calculated energies, instead of  $E^{\text{vdW,DFT}}$ , may also be used as target values, *e.g.* the total DFT energy  $E^{\text{tot,DFT}}$ . In that case, the response entering the response vector in Equation (5.42) would be a ‘‘hypothetical’’ vdW contribution  $\tilde{E}^{\text{vdW,DFT}}$  derived from the calculated total DFT energy  $E^{\text{tot,DFT}}$ , *i.e.*

$$\tilde{E}^{\text{vdW,DFT}} = E^{\text{tot,DFT}} - E^{\text{Coulomb,FF}} - E^{\text{tors,FF}} - E^{\text{angles,FF}} - E^{\text{bonds,FF}}, \quad (5.45)$$

where the individual FF contributions are given in Equations (5.7), (5.6), (5.5), and (5.4).

The torsions contribution to the potential energy is given in Equation (5.6). Using an equivalent approach as before, one may estimate the parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  by again making use of the different regression models laid out in Subsection 5.2.2. To that end, the calculated total DFT energy  $E^{\text{tot,DFT}}$  may again serve as target data. The response vector in Equation (5.12) is written as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{\tilde{i}} \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \tilde{E}_1^{\text{torsions,DFT}} \\ \vdots \\ \tilde{E}_{\tilde{i}}^{\text{torsions,DFT}} \\ \vdots \\ \tilde{E}_n^{\text{torsions,DFT}} \end{pmatrix}, \quad (5.46)$$

## Chapter 5. Force Field Parameterization Using Regularized Linear Regression

where  $\tilde{E}_{\tilde{i}}^{\text{torsions,DFT}}$  denotes the “hypothetical” torsions contribution of conformer  $\tilde{i}$  derived from the calculated total DFT energy  $E^{\text{tot,DFT}}$ , *i.e.*

$$\tilde{E}_{\tilde{i}}^{\text{torsions,DFT}} = E^{\text{tot,DFT}} - E^{\text{Coulomb,FF}} - E^{\text{vdW,FF}} - E^{\text{angles,FF}} - E^{\text{bonds,FF}}, \quad (5.47)$$

where the individual FF contributions are given in Equations (5.7), (5.8), (5.5), and (5.4). The vector  $\boldsymbol{\beta}$  containing the regression coefficients given in Equation (5.13) is written as

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} E^{\text{offset}} \\ V_1^{11} \\ V_2^{11} \\ V_3^{11} \\ V_1^{12} \\ V_2^{12} \\ V_3^{12} \\ \vdots \end{pmatrix}, \quad (5.48)$$

*i.e.* the vector  $\boldsymbol{\beta}$  contains all possible  $V_m^{ij}$ ,  $m = 1, 2, 3$ , parameters as regression coefficients attributed to their individual pairs of classes of atoms within the description of the OPLS-AA FF, and the “intercept”  $\beta_0$  is just an arbitrary energetic potential offset  $E^{\text{offset}}$ . Taking into account the formulation of the torsional contribution to the potential energy in the OPLS-AA FF description given in Equation (5.6), the vector  $\mathbf{x}_{\tilde{i}}$  containing the predictor variables (see Equation (5.10)) is immediately written as

$$\mathbf{x}_{\tilde{i}} = \begin{pmatrix} 1 \\ x_{\tilde{i}1} \\ x_{\tilde{i}2} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 \\ \left( \sum_{11} \frac{1}{2} (1 + \cos(\phi^{11})) \right)_{\tilde{i}} \\ \left( \sum_{11} \frac{1}{2} (1 - \cos(2\phi^{11})) \right)_{\tilde{i}} \\ \left( \sum_{11} \frac{1}{2} (1 + \cos(3\phi^{11})) \right)_{\tilde{i}} \\ \left( \sum_{12} \frac{1}{2} (1 + \cos(\phi^{12})) \right)_{\tilde{i}} \\ \left( \sum_{12} \frac{1}{2} (1 - \cos(2\phi^{12})) \right)_{\tilde{i}} \\ \left( \sum_{12} \frac{1}{2} (1 + \cos(3\phi^{12})) \right)_{\tilde{i}} \\ \vdots \end{pmatrix}, \quad (5.49)$$

where the sums indicate summations over all pairs of 1-4 atoms with the same individual atom classes for all four atoms involved in the torsion. As before, using these definitions the approach for estimating the torsions parameters  $V_m^{ij}$ ,  $m = 1, 2, 3$ , using different regression models is straightforward and has already been laid out in Subsection 5.2.2 where the objective functions to minimize for conventional multiple linear regression, Ridge regression, and the LASSO have been given in Equations (5.17), (5.23), and (5.33), respectively.

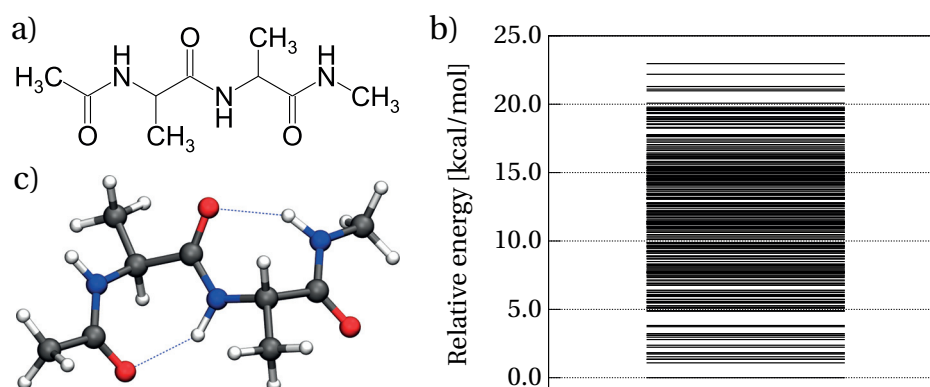


Figure 5.2 – (a) Structural formula of AcAla<sub>2</sub>NMe. (b) Energy hierarchy of conformers calculated at the PBE+vdW<sup>TS</sup> DFT level. (c) The conformer with lowest energy is illustrated.

### FFAFFURR: Output

After having adjusted the FF parameters of choice using different methods described above, a TINKER-specific potential energy parameter file named `opl_saa-ffaffurr.prm` is written out and may be immediately used for further calculations.

## 5.3 Results

A *proof of principle* of the method is intended using the toy model of AcAla<sub>2</sub>NMe in the gas phase. The structural formula of the peptide is provided in Figure 5.2(a).

In order to yield a set of conformers that may be used as training or test data, a conformational search algorithm equivalent to the one laid out in Section 3.3 is undergone. First, a global conformational search explained in Subsection 2.3.1 is performed at the FF level using the original OPLS-AA FF formulation and parameters. To that end, the basin-hopping approach described in Section 2.4 is applied using the `scan` program of the TINKER program. All torsional modes are thereby taken into consideration and default search parameters are used, *i.e.* an energy threshold for local minima of 100 kcal/mol and a convergence criterion for local geometry optimizations of 0.0001 kcal/mol·Å. In total 311 conformers were found. All conformers are then geometry optimized at the DFT level using FHI-aims, more precisely at the PBE+vdW<sup>TS</sup> level using `tier 1` basis sets and `light` settings. Relaxation is accomplished using a trust radius method version of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm [338]. After convergence, a clustering scheme is applied using simple root-mean-square deviations (RMSD) of atomic positions in order to rule out duplicates. Hierarchical clustering is thereby achieved by applying the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [340] method implemented in Python's SciPy [341] library. Following that, further relaxation is accomplished at the PBE+vdW<sup>TS</sup> level using `tier 2` basis sets and `tight` settings. After clustering, this results in 231 conformers. The corresponding energy hierarchy and a depiction of the lowest-energy conformer are shown in Figures 5.2(b) and (c), respectively.

## Chapter 5. Force Field Parameterization Using Regularized Linear Regression

Table 5.1 – Summary of chemical symbols, TINKER-specific symbols, atom types, and atom classes of the atoms of the system of AcAla<sub>2</sub>NMe using the standard notation of the OPLS-AA FF file distributed with the TINKER package.

Atom number	Chemical symbol	TINKER-specific symbol	Atom type	Atom class
1	C	CT	80	13
2	C	C	177	3
3	O	O	178	4
4	H	HC	85	46
5	H	HC	85	46
6	H	HC	85	46
7	N	N	180	24
8	C	CT	166	13
9	C	C	177	3
10	O	O	178	4
11	H	H	183	45
12	H	HC	85	46
13	C	CT	80	13
14	H	HC	85	46
15	H	HC	85	46
16	H	HC	85	46
17	N	N	180	24
18	C	CT	166	13
19	C	C	177	3
20	O	O	178	4
21	H	H	183	45
22	H	HC	85	46
23	C	CT	80	13
24	H	HC	85	46
25	H	HC	85	46
26	H	HC	85	46
27	N	N	180	24
28	C	CT	184	13
29	H	H	183	45
30	H	HC	85	46
31	H	HC	85	46
32	H	HC	85	46

Out of the 231 conformers, half of them, *i.e.* 115, were selected at random for the set of training data, leaving 116 conformers for the set of test data. In order to compare the energetic performances between calculations at the FF level and the DFT level mean absolute errors (MAEs) and maximum errors (MEs) are again used as a quality measure, as explained in Subsection 4.2.3. Using the test set of conformers and the original parameters of the OPLS-AA FF that are distributed with the TINKER package yields a MAE of 2.55 kcal/mol and a ME of 10.45 kcal/mol when compared to the PBE+vdW<sup>TS</sup> level, which is in accordance to what one would expect when taking into considerations the findings in Chapter 4. Table 5.1 summarizes chemical symbols, TINKER-specific symbols, atom types, and atom classes of the atoms of AcAla<sub>2</sub>NMe using the standard notation of the OPLS-AA FF file distributed with the TINKER package. The atom types are illustrated for their respective atoms in Figure 5.3.

Because all input structures of the training set (as well as the test set) were geometry optimized beforehand at the PBE+vdW<sup>TS</sup> level as explained above, the equilibrium parameters  $r_{ij}^0$



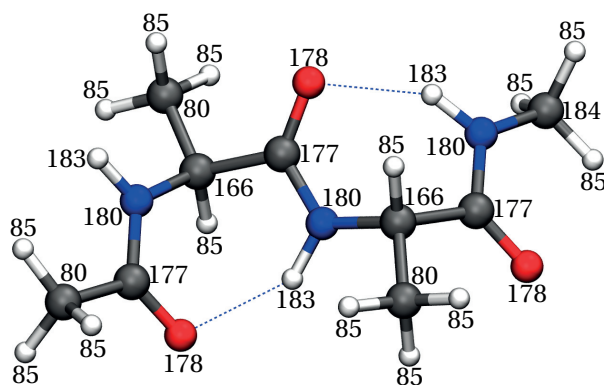


Figure 5.3 – Illustration of the atom types according to the standard notation of the OPLS-AA FF file distributed with the TINKER package for the system of the system of AcAla<sub>2</sub>NMe.

and  $\theta_{ij}^0$  between pairs or triplets of atoms are adjusted by averaging over all corresponding atomic pairwise distances and angles of the same pair or triplet of atom classes over all input files, as explained in Subsection 5.2.3. Hence, the energetic contributions from the bonds and angles contributions in Equations (5.4) and (5.5) should diminish as the minima of the terms of quadratic form are shifted towards the average of the “real” distances and angles. Indeed, already this simple adjustment yields a decrease in the MAE by 0.75 kcal/mol to just 1.81 kcal/mol and a ME of 8.93 kcal/mol for the test set. It is interesting to note that the actual values of  $r_{ij}^0$  are not altered by more than 0.03 Å, and the actual values of  $\theta_{ij}^0$  are not altered by more than 5.3°. Adjusting only  $r_{ij}^0$  in the same manner yields a MAE of 2.25 kcal/mol and a ME of 10.00 kcal/mol. Modifying only  $\theta_{ij}^0$  in the same manner yields a MAE of 1.88 kcal/mol and a ME of 9.26 kcal/mol. This strongly indicates that the quadratic form of the corresponding potential energy terms is generally not suitable to accurately describe energetics of conformers that do not lie in a minimum on the PES, especially considering the “angles” contributions. A summary of all values of empirical parameters  $r_{ij}^0$  and  $\theta_{ij}^0$  is provided in Table 5.2. The adjusted values of  $r_{ij}^0$  and  $\theta_{ij}^0$  will be used going forward.

Estimating the empirical partial charge parameters  $q_i$  from the training set as explained in Subsection 5.2.3 using Hirshfeld charges and ESP charges yields a MAE of 2.62 kcal/mol (ME of 10.02 kcal/mol) and a MAE of 2.93 kcal/mol (ME of 10.17 kcal/mol), respectively. A summary of all values of empirical parameters  $q_i$  is provided in Table 5.3. The large error deviations indicate that the new charge parameter estimates are clearly not suitable to accurately describe the conformational energy hierarchy of the test set. The reason for that is not clear, but improvement could be achieved in some studies by simply applying scaling factors for the electrostatic interactions [403]. Figure 5.4 shows obtained MAEs and MEs when multiplying the obtained Hirshfeld and ESP partial charges with scaling factors from 0.50 to 2.00. Interestingly, the “optimal” scaling factor for Hirshfeld charges is found to be 1.55 resulting in a MAE of 2.00 kcal/mol while the “optimal” scaling factor for ESP charges is found to be 0.75 resulting in a MAE of 2.12 kcal/mol. However, performance is still significantly worse in comparison to using the partial charges of the original OPLS-AA FF. The distributions of Hirshfeld and

## Chapter 5. Force Field Parameterization Using Regularized Linear Regression

Table 5.2 – Original OPLS-AA FF parameters  $r_{ij}^0$  and  $\theta_{ij}^0$  and their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by averaging over all corresponding atomic pairwise distances and angles of the same pair or triplet of atom classes over all input files.

Atom class pair	$r_{ij}^0$ [Å] (original FF)	$r_{ij}^0$ [Å] (adjusted FF)	Atom class triplet	$\theta_{ij}^0$ [°] (original FF)	$\theta_{ij}^0$ [°] (adjusted FF)
(3, 4)	1.2290	1.2298	(4, 3, 13)	120.4000	120.8596
(3, 13)	1.5220	1.5374	(4, 3, 24)	122.9000	121.8180
(3, 24)	1.3350	1.3681	(13, 3, 24)	116.6000	117.2861
(13, 13)	1.5290	1.5329	(3, 13, 13)	111.1000	111.8153
(13, 24)	1.4490	1.4590	(3, 13, 24)	110.1000	111.2413
(13, 46)	1.0900	1.0975	(13, 13, 24)	109.7000	112.3284
(24, 45)	1.0100	1.0167	(3, 13, 46)	109.5000	108.5675
			(13, 13, 46)	110.7000	109.9970
			(24, 13, 46)	109.5000	108.9251
			(46, 13, 46)	107.8000	108.4278
			(3, 24, 13)	121.9000	127.2405
			(3, 24, 45)	119.8000	114.6307
			(13, 24, 45)	118.4000	116.7539

Table 5.3 – Original OPLS-AA FF parameters  $q_i$  and their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by averaging over all Hirshfeld or ESP charges of the same atom type over all input files.

Atom type	$q_i$ (original FF)	$q_i$ (Hirshfeld) (adjusted FF)	$q_i$ (ESP) (adjusted FF)
80	-0.1800	-0.1143	-0.5650
85	0.0600	0.0461	0.1257
166	0.1400	0.0156	0.4510
177	0.5000	0.1487	0.4447
178	-0.5000	-0.2701	-0.5127
180	-0.5000	-0.0928	-0.4344
183	0.3000	0.1218	0.2734
184	0.0200	-0.0559	-0.2803

ESP charges over the training set of conformers are presented in Figure 5.5. While Hirshfeld charges are rather well-defined over all atom types of the system, the same statement does not hold true concerning ESP charges of “buried” atoms, *i.e.* saturated carbons (atom types 80, 166, 177, 184) and nitrogens (atom type 180), which indicates a drawback of the method that has already been discussed elsewhere [404–406]. Going forward, the empirical partial charge parameters of the original OPLS-AA FF will be used.

Estimating the empirical vdW parameters  $\sigma_{ij}$  by using the atomic Hirshfeld partitioning scheme as explained in Subsection 5.2.3 yields a decrease in the MAE by 0.04 kcal/mol to 1.77 kcal/mol and a ME of 5.91 kcal/mol for the test set. A summary of all values of empirical

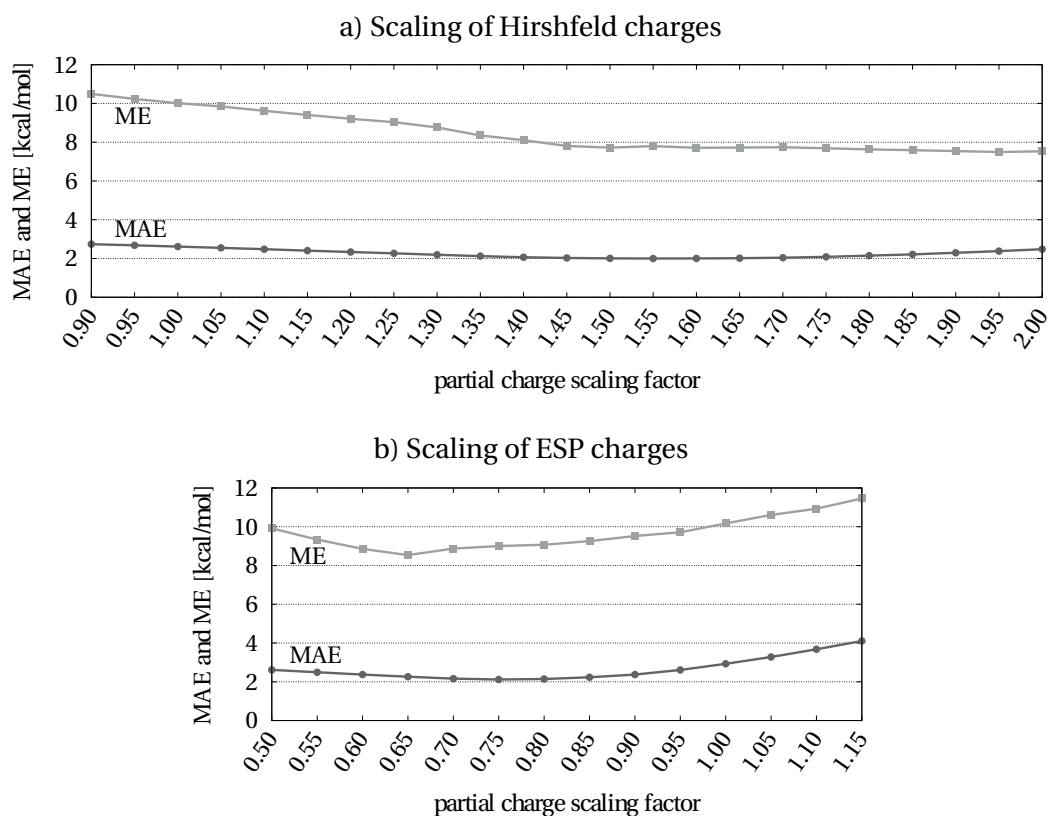


Figure 5.4 – Obtained MAEs (dark-gray) and MEs (light-gray) for the test set of AcAla<sub>2</sub>NMe when multiplying the obtained Hirshfeld and ESP partial charges with scaling factors.

parameters  $\sigma_{ij}$  is provided in Table 5.4.

Because the energetic dispersion corrections are quantitatively small when compared to the total DFT energy of a system of small size like AcAla<sub>2</sub>NMe, one must expect only small improvements in the energetic description of the FF when estimating empirical vdW parameters  $\epsilon_{ij}$  by using regression models with *a posteriori* calculated vdW contributions as target data, as explained in Subsection 5.2.3. Applying conventional multiple linear regression with the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> energy  $E^{\text{vdW}^{\text{TS}},\text{DFT}}$  taken as response data, one further decreases the MAE by 0.05 kcal/mol to 1.72 kcal/mol and yields a ME of 5.54 kcal/mol for the test set. A summary of all values of empirical parameters  $\epsilon_{ij}$  is provided in Table 5.5. Because no restrictions have been made for the empirical parameters  $\epsilon_{ij}$ , a few estimated parameters become negative (*e.g.* for atom type pair (166,166) or (180,184)) which obviously contradicts the physical nature of the Lennard-Jones 12-6 potential that requires positive values of  $\epsilon_{ij}$ . In addition, some values become rather large when compared to their counterparts of the original OPLS-AA, see *e.g.* for atom type pair (80,80). Because one might suspect overdetermination due to the large number of predictor variables (the  $\epsilon_{ij}$ 's) LASSO regression instead of conventional multiple linear regression is laid out with  $E^{\text{vdW}^{\text{TS}},\text{DFT}}$  again taken as response data. To that end, 50 logarithmically equidistantly distributed values in the range of 0.0001

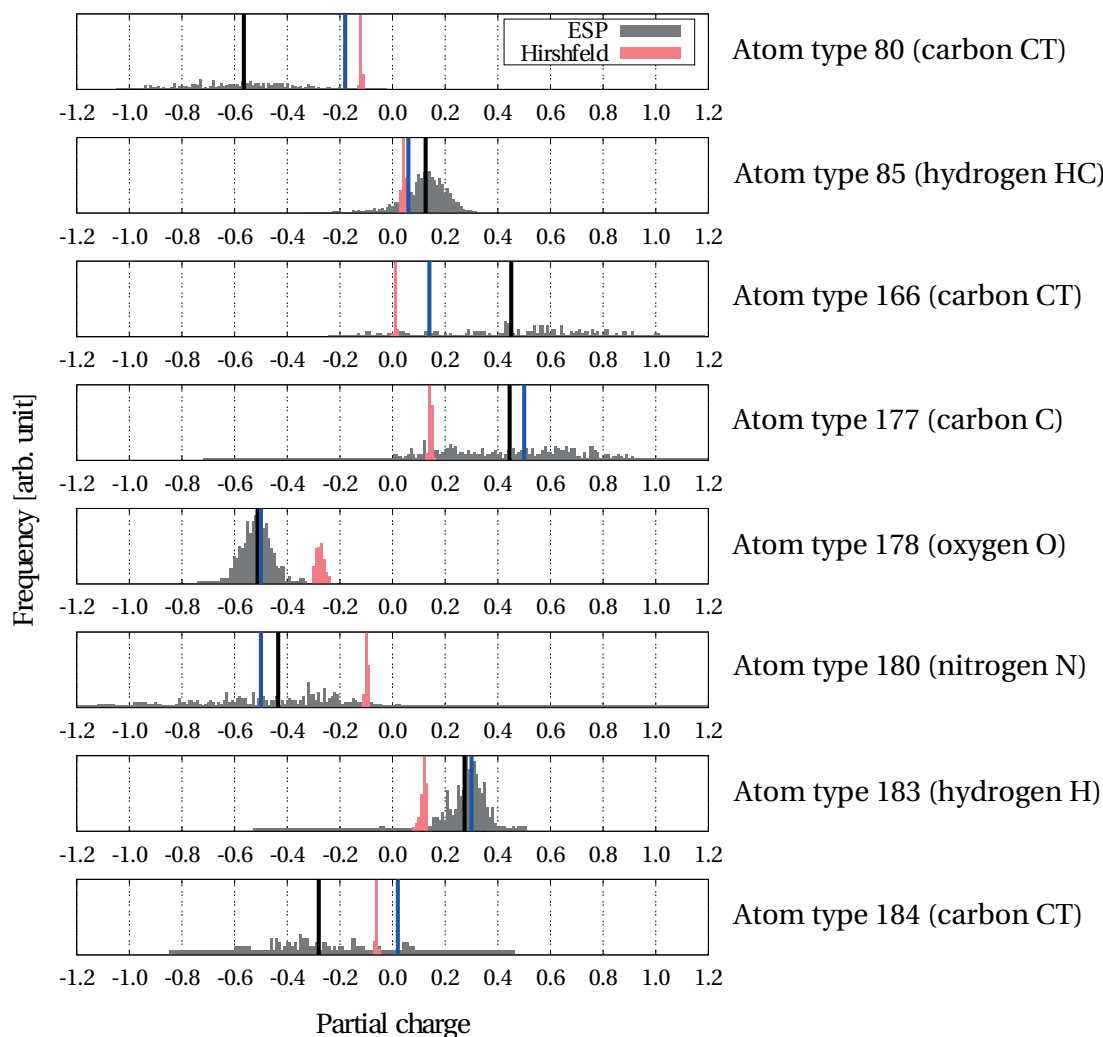


Figure 5.5 – Distributions of Hirshfeld (red) and ESP (gray) charges over the training set of conformers for the system of AcAla<sub>2</sub>NMe. For visibility purposes, the frequencies of the ESP charges have been scaled by a factor of 3. Vertical blue lines denote partial charge parameters of the original OPLS-AA FF while vertical black lines denote the average of the ESP values (compare with Table 5.3).

to 4 have been applied for the regularization parameter  $\lambda$ . Figure 5.6 shows the respectively obtained regression coefficients  $\epsilon_{ij}$ , the residual sum of squares  $RSS$  (see Equation (5.21)), and calculated MAEs and MEs for the test set. As expected, the MAE for the test set is rather unperturbed due to the small influence that the dispersion correction contributes to the total energy of a system of such small size. For large  $\lambda$ , the bias is so large that the estimates shrink to zero. For  $\lambda \rightarrow 0$ , the bias vanishes and the limit of conventional multiple linear regression is yielded with results described above. An “optimal” value of  $\lambda$  could be considered to lie in the range  $0.01 \lesssim \lambda \lesssim 0.1$ . Indeed, for  $\lambda = 0.018$  and restricting the regression coefficients  $\epsilon_{ij}$  to non-negative values, the ME for the test set decreases by 0.37 kcal/mol to 5.17 kcal/mol,

and a MAE of 1.69 kcal/mol is yielded. A summary of all corresponding values of empirical parameters  $\epsilon_{ij}$  is provided in Table 5.6.

Estimating the torsions empirical parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  from the training set using conventional multiple linear regression with the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data, as explained in Subsection 5.2.3, yields a MAE of 0.88 kcal/mol and a ME of 3.20 kcal/mol. A summary of values of empirical parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  is provided in Table 5.7. Note that only those parameters have been taken into account that were already non-zero in the original OPLS-AA FF description. Some estimated values are rather large (*e.g.* for atom class quadruplet (13,3,24,13) or (4,3,24,13)), thus hinting at a *ill-conditioned* solution. Hence, Ridge regression instead of conventional multiple linear regression is laid out with the total DFT energy again taken as response data. To that end, 50 logarithmically equidistantly distributed values in the range of 0.0001 to 4 have been again applied for the regularization parameter  $\lambda$ . Figure 5.7 shows the respectively obtained regression coefficients  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$ , the residual sum of squares *RSS* (see Equation (5.21)), and calculated MAEs and MEs for the test set. Observing the smooth behavior of the regression coefficients with respect to varying the regularization parameter  $\lambda$ , it is clear that the solution is not *ill-conditioned*. Applying a regularization here just means to “artificially” shrink the torsional parameter estimates towards the origin, thus resulting in no further improvement of the energetic FF description, as seen by the nearly constant MAEs and MEs for the test set over the whole  $\lambda$  range. Taking into consideration the rather arbitrary restriction that until now only those parameters have been taken into account that were already non-zero in the original OPLS-AA FF description, one may repeat the procedure but estimating *all* possible parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  using conventional multiple linear regression with the total DFT energy taken as response data. Doing so yields a MAE of 0.75 kcal/mol and a ME of 2.87 kcal/mol. A summary of values of empirical parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  is provided in Table 5.8. While the MAE for the test set is well within “chemical accuracy” of 1 kcal/mol, many estimated parameters become very large, *e.g.* for the atom class quadruplet (4,3,24,45) or (13,3,24,45). Due to the large number of predictor variables and the possible likelihood of overdetermination, the LASSO regression model appears to be appealing for the problem in question. Indeed, already for regularization parameters  $\lambda \gtrsim 0.001$  the solution is significantly stabilized without sacrificing energetic performance as seen by the nearly unfazed MAE for the test set presented in Figure 5.8, where again obtained regression coefficients  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$ , the residual sum of squares *RSS*, and calculated MAEs and MEs for the test set are shown. *E.g.* for  $\lambda = 0.0017$ , a MAE of 0.76 kcal/mol and a ME of 2.81 kcal/mol is yielded. A summary of the corresponding empirical parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  is provided in Table 5.9.

In summary, by estimating different empirical FF parameters using different adjustment models, it is possible to reach a mean energetic description of the test set of hierarchical minima on the PES for the simple system of AcAla<sub>2</sub>NMe well within “chemical accuracy” of 1 kcal/mol. The prime reason for this approach to find success for this simple system is found in the “flexibility” of the torsional contributions to the energetic FF description in terms of their predictive ability to describe energetic changes in the molecule, hence why the approach

## Chapter 5. Force Field Parameterization Using Regularized Linear Regression

Table 5.4 – Original OPLS-AA FF parameters  $\sigma_{ij}$  and their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by the atomic Hirshfeld partitioning scheme as explained in Subsection 5.2.3.

Atom type pair	$\sigma_{ij}$ [Å] (original FF)	$\sigma_{ij}$ [Å] (adjusted FF)	Atom type pair	$\sigma_{ij}$ [Å] (original FF)	$\sigma_{ij}$ [Å] (adjusted FF)
(80, 80)	3.5000	2.9424	(166, 177)	3.6228	3.0162
(80, 85)	2.9580	2.6568	(166, 178)	3.2187	2.9086
(80, 166)	3.5000	2.9758	(166, 180)	3.3727	2.9126
(80, 177)	3.6228	2.9828	(166, 184)	3.5000	2.9632
(80, 178)	3.2187	2.8752	(177, 177)	3.7500	3.0233
(80, 180)	3.3727	2.8792	(177, 178)	3.3317	2.9157
(80, 184)	3.5000	2.9297	(177, 180)	3.4911	2.9196
(85, 85)	2.5000	2.3712	(177, 184)	3.6228	2.9702
(85, 166)	2.9580	2.6902	(178, 178)	2.9600	2.8081
(85, 177)	3.0619	2.6972	(178, 180)	3.1016	2.8120
(85, 178)	2.7203	2.5896	(178, 184)	3.2187	2.8626
(85, 180)	2.8504	2.5936	(180, 180)	3.2500	2.8159
(85, 184)	2.9580	2.6441	(180, 184)	3.3727	2.8665
(166, 166)	3.5000	3.0092			

Table 5.5 – Original OPLS-AA FF parameters  $\epsilon_{ij}$  and their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by using conventional multiple linear regression with the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> energy  $E^{\text{vdW}^{\text{TS}},\text{DFT}}$  taken as response data, as explained in Subsection 5.2.3.

Atom type pair	$\epsilon_{ij}$ [kcal/mol] (original FF)	$\epsilon_{ij}$ [kcal/mol] (adjusted FF)	Atom type pair	$\epsilon_{ij}$ [kcal/mol] (original FF)	$\epsilon_{ij}$ [kcal/mol] (adjusted FF)
(80, 80)	0.0660	1.0356	(166, 177)	0.0832	0.4336
(80, 85)	0.0445	-0.0083	(166, 178)	0.1177	0.0556
(80, 166)	0.0660	0.0626	(166, 180)	0.1059	0.2985
(80, 177)	0.0832	0.4963	(166, 184)	0.0660	0.0384
(80, 178)	0.1177	0.1748	(177, 177)	0.1050	0.2167
(80, 180)	0.1059	0.3425	(177, 178)	0.1485	0.0160
(80, 184)	0.0660	0.5769	(177, 180)	0.1336	0.1229
(85, 85)	0.0300	0.0005	(177, 184)	0.0832	0.8924
(85, 166)	0.0445	0.0161	(178, 178)	0.2100	0.2931
(85, 177)	0.0561	0.0072	(178, 180)	0.1889	0.1011
(85, 178)	0.0794	0.0085	(178, 184)	0.1177	0.1389
(85, 180)	0.0714	-0.0009	(180, 180)	0.1700	0.1517
(85, 184)	0.0445	0.1027	(180, 184)	0.1059	-0.0683
(166, 166)	0.0660	-0.2122			

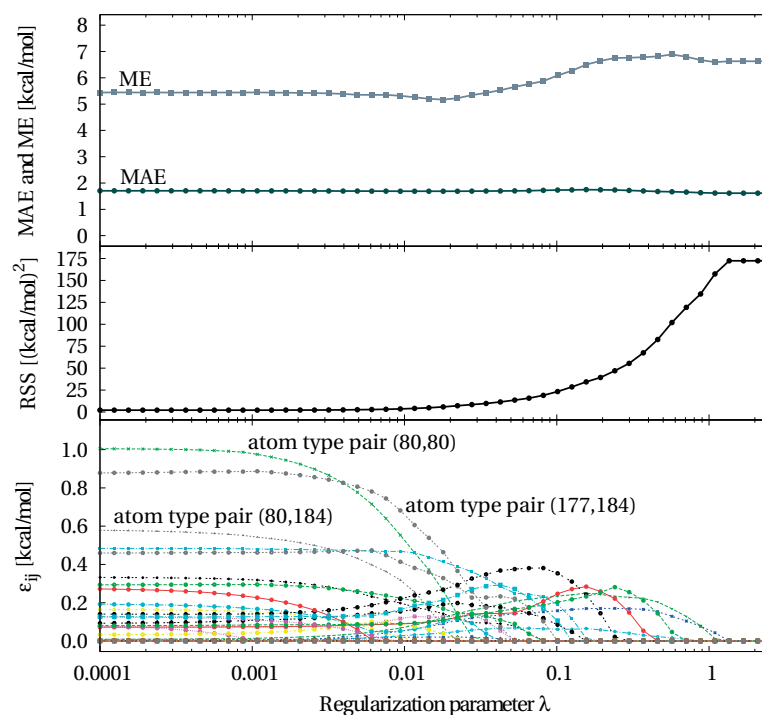


Figure 5.6 – Estimated regression coefficients  $\epsilon_{ij}$  (bottom), the residual sum of squares  $RSS$  (middle), and calculated MAEs and MEs (top) for the test set of AcAla<sub>2</sub>NMe obtained by using LASSO regression with the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> energy  $E^{\text{vdW}^{\text{TS}},\text{DFT}}$  taken as response data.

Table 5.6 – Original OPLS-AA FF parameters  $\epsilon_{ij}$  and their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by using LASSO regression ( $\lambda = 0.018$ ) with the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> energy  $E^{\text{vdW}^{\text{TS}},\text{DFT}}$  taken as response data, as explained in Subsection 5.2.3.

Atom type pair	$\epsilon_{ij}$ [kcal/mol] (original FF)	$\epsilon_{ij}$ [kcal/mol] (adjusted FF)	Atom type pair	$\epsilon_{ij}$ [kcal/mol] (original FF)	$\epsilon_{ij}$ [kcal/mol] (adjusted FF)
(80, 80)	0.0660	0.2618	(166, 177)	0.0832	0.3289
(80, 85)	0.0445	0.0912	(166, 178)	0.1177	0.1094
(80, 166)	0.0660	0.1253	(166, 180)	0.1059	0.1687
(80, 177)	0.0832	0.4133	(166, 184)	0.0660	0.0000
(80, 178)	0.1177	0.0000	(177, 177)	0.1050	0.1012
(80, 180)	0.1059	0.1424	(177, 178)	0.1485	0.0382
(80, 184)	0.0660	0.0697	(177, 180)	0.1336	0.2685
(85, 85)	0.0300	0.0000	(177, 184)	0.0832	0.4638
(85, 166)	0.0445	0.1223	(178, 178)	0.2100	0.0000
(85, 177)	0.0561	0.0206	(178, 180)	0.1889	0.1103
(85, 178)	0.0794	0.0000	(178, 184)	0.1177	0.0000
(85, 180)	0.0714	0.0420	(180, 180)	0.1700	0.2140
(85, 184)	0.0445	0.1912	(180, 184)	0.1059	0.0000
(166, 166)	0.0660	0.0000			

## Chapter 5. Force Field Parameterization Using Regularized Linear Regression

Table 5.7 – Original OPLS-AA FF parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  as well as their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by using conventional multiple linear regression with the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data, as explained in Subsection 5.2.3. Here, only those parameters have been taken into account that were already non-zero in the original OPLS-AA FF description.

Atom class quadruplet	$V_1^{ij}$ [kcal/mol] (original FF)	$V_1^{ij}$ [kcal/mol] (adjusted FF)	$V_2^{ij}$ [kcal/mol] (original FF)	$V_2^{ij}$ [kcal/mol] (adjusted FF)	$V_3^{ij}$ [kcal/mol] (original FF)	$V_3^{ij}$ [kcal/mol] (adjusted FF)
(24,3,13,13)	1.173	2.378	0.189	0.001	-2.200	-1.463
(24,3,13,24)	1.816	1.673	1.222	2.681	1.581	0.721
(13,3,24,13)	2.300	2.964	6.089	-12.728	—	—
(3,13,24,3)	-2.365	1.146	0.912	-0.771	-0.850	0.026
(4,3,24,13)	—	—	6.089	15.589	—	—
(4,3,24,45)	—	—	4.900	-3.565	—	—
(13,3,24,45)	—	—	4.900	-2.884	—	—
(13,13,24,3)	—	—	0.462	0.573	—	—
(3,13,13,46)	—	—	—	—	-0.100	3.542
(24,13,13,46)	—	—	—	—	0.464	8.580
(46,13,13,46)	—	—	—	—	0.300	-9.084

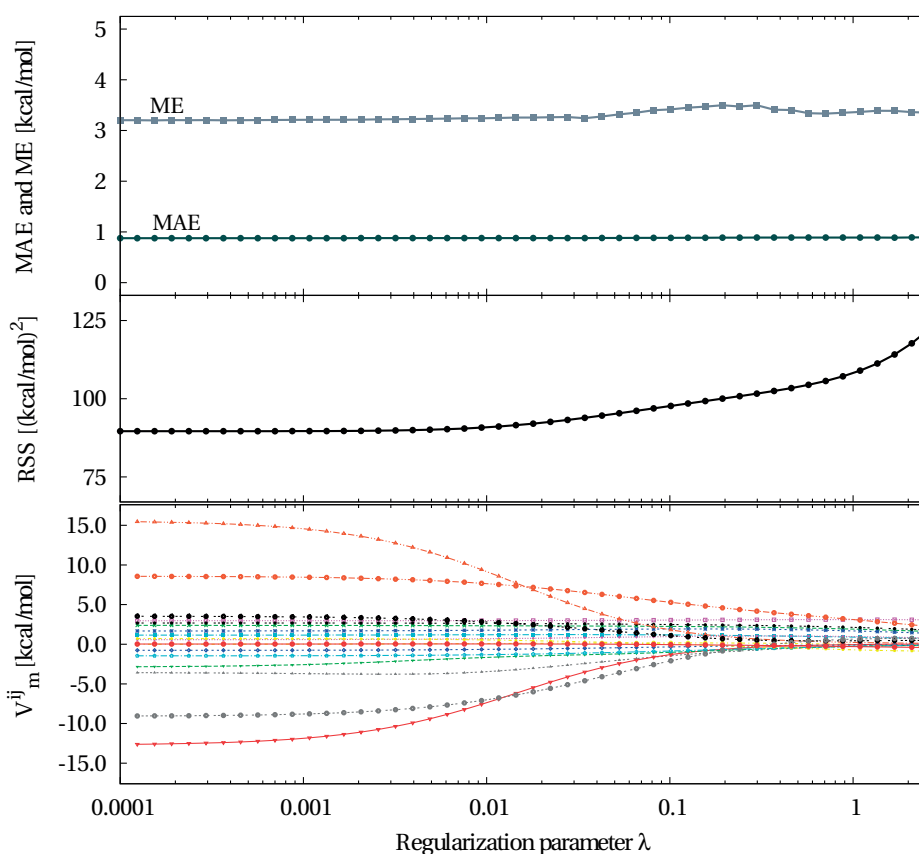


Figure 5.7 – Estimated regression coefficients  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  (bottom), the residual sum of squares  $RSS$  (middle), and calculated MAEs and MEs (top) for the test set of AcAla<sub>2</sub>NMe obtained by using Ridge regression with the total DFT energy taken as response data. Here, only those parameters have been taken into account that were already non-zero in the original OPLS-AA FF description.



Table 5.8 – Original OPLS-AA FF parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  as well as their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by using conventional multiple linear regression with the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data, as explained in Subsection 5.2.3.

Atom class quadruplet	$V_1^{ij}$ [kcal/mol] (original FF)	$V_1^{ij}$ [kcal/mol] (adjusted FF)	$V_2^{ij}$ [kcal/mol] (original FF)	$V_2^{ij}$ [kcal/mol] (adjusted FF)	$V_3^{ij}$ [kcal/mol] (original FF)	$V_3^{ij}$ [kcal/mol] (adjusted FF)
(24,3,13,13)	1.173	1.648	0.189	-0.269	-1.200	-1.476
(24,3,13,24)	1.816	1.529	1.222	1.739	1.581	0.883
(4,3,24,13)	0.000	148.170	6.089	17.607	0.000	-19.083
(4,3,24,45)	0.000	-1942.069	4.900	11.371	0.000	229.955
(13,3,24,13)	2.300	-24.370	6.089	-10.570	0.000	3.955
(13,3,24,45)	0.000	-2120.841	4.900	-11.824	0.000	256.181
(3,13,13,46)	0.000	781.300	0.000	-220.433	-0.100	3.770
(24,13,13,46)	0.000	847.011	0.000	-162.283	0.464	5.135
(46,13,13,46)	0.000	758.724	0.000	-245.675	0.300	-9.523
(3,13,24,3)	-2.365	1.596	0.912	-1.145	-0.850	1.321
(13,13,24,3)	0.000	1.241	0.462	0.380	0.000	-1.059

of using multiple linear regression or LASSO regression for estimating empirical torsions parameters works rather well. However, the main drawback of the overall approach consists in the poor energetic hierarchies obtained when estimating empirical partial charges from the Hirshfeld partitioning scheme or by fitting the partial charges to reproduce the electrostatic potential (ESP). This flaw is especially concerning for systems for which the energetic FF description is dominated by Coulomb interaction contributions like systems involving a metal cation.

In order to quantify such differences in the energetic description of peptide-cation systems, the same study that has been presented here is repeated for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup>. The conformational search algorithm equivalent to the one laid out above results in a total of 327 conformers at the PBE+vdW<sup>TS</sup> DFT level. Selecting half of them, *i.e.* 163, at random for the set of training data leaves 164 conformers for the set of test data. Using the test set of conformers and the original parameters of the OPLS-AA FF that are distributed with the TINKER package yields a MAE of 4.08 kcal/mol and a ME of 19.82 kcal/mol when compared to the PBE+vdW<sup>TS</sup> level, which is a significantly worse performance when compared to the values of 2.55 kcal/mol and 10.45 kcal/mol for the bare system of AcAla<sub>2</sub>NMe, as one would expect when taking into considerations the findings in Chapter 4. One thereby needs to keep in mind that the only difference consists in an additional Na<sup>+</sup> cation (atom type 349 and atom class 69 within the nomenclature of TINKER) of empirical partial charge  $q_{\text{Na}^+} = +1$  that interacts with other atoms within the OPLS-AA FF description exclusively by means of Coulomb and vdW interaction, see Equations (5.7) and (5.8), respectively. And yet, this simple change alone results in a much worse energetic performance of the FF, as the MAE for the test set increases by 1.53 kcal/mol and the corresponding ME increases by 9.37 kcal/mol with respect to the bare peptide system. Hence, the current procedure cannot be expected to fully “compensate” this discrepancy in energetic performance, which is because it mainly relies on adjusting the torsional contributions that do not include the added Na<sup>+</sup> in any form within the OPLS-AA FF

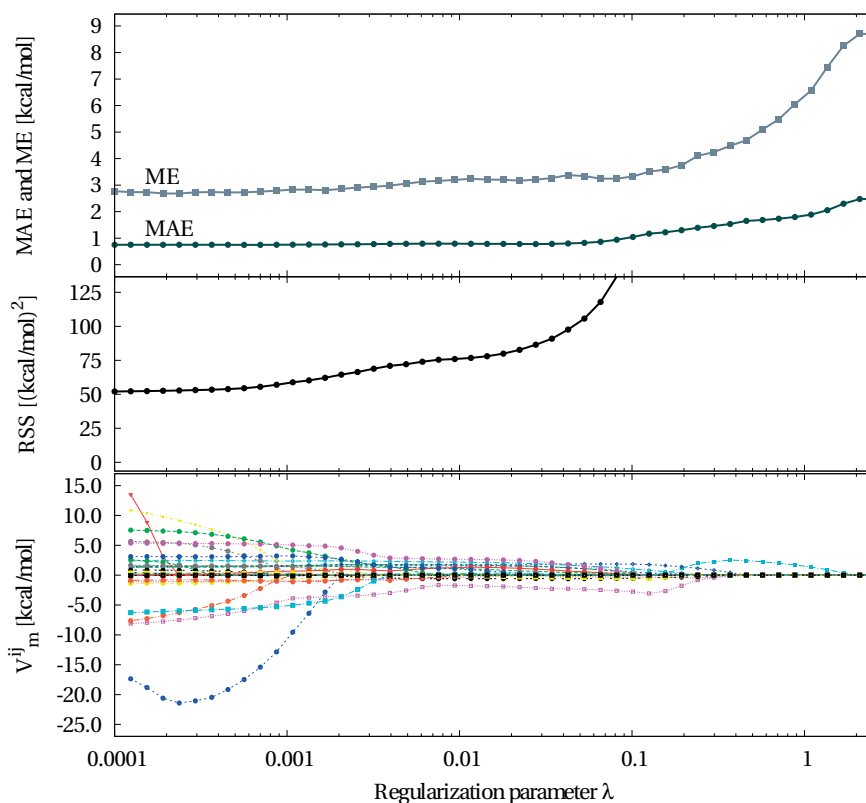


Figure 5.8 – Estimated regression coefficients  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  (bottom), the residual sum of squares  $RSS$  (middle), and calculated MAEs and MEs (top) for the test set of AcAla<sub>2</sub>NMe obtained by using LASSO regression with the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data.

Table 5.9 – Original OPLS-AA FF parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  as well as their adjusted counterparts for the system of AcAla<sub>2</sub>NMe obtained by using LASSO regression ( $\lambda = 0.0017$ ) with the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data, as explained in Subsection 5.2.3.

Atom class quadruplet	$V_1^{ij}$ [kcal/mol] (original FF)	$V_1^{ij}$ [kcal/mol] (adjusted FF)	$V_2^{ij}$ [kcal/mol] (original FF)	$V_2^{ij}$ [kcal/mol] (adjusted FF)	$V_3^{ij}$ [kcal/mol] (original FF)	$V_3^{ij}$ [kcal/mol] (adjusted FF)
(24,3,13,13)	1.173	1.682	0.189	0.070	-1.200	-0.478
(24,3,13,24)	1.816	1.520	1.222	2.394	1.581	-0.028
(4,3,24,13)	0.000	-3.608	6.089	1.091	0.000	-0.988
(4,3,24,45)	0.000	0.118	4.900	0.000	0.000	0.759
(13,3,24,13)	2.300	0.000	6.089	0.000	0.000	0.808
(13,3,24,45)	0.000	0.000	4.900	0.000	0.000	3.204
(3,13,13,46)	0.000	0.000	0.000	0.000	-0.100	2.940
(24,13,13,46)	0.000	0.000	0.000	0.000	0.464	4.870
(46,13,13,46)	0.000	0.000	0.000	-2.843	0.300	-4.347
(3,13,24,3)	-2.365	1.511	0.912	-0.920	-0.850	0.179
(13,13,24,3)	0.000	1.731	0.462	0.092	0.000	0.000

description. Nevertheless, an attempt is undergone and key results are briefly summarized in the following. A detailed listing of all values of adjusted empirical parameters as well as plotted regression coefficients, the *RSS*, and calculated MAEs and MEs for the test set obtained by using different regression models is provided in Appendix A.

Adjusting the equilibrium parameters  $r_{ij}^0$  and  $\theta_{ij}^0$  between pairs or triplets of atoms by averaging over all corresponding atomic pairwise distances and angles of the same pair or triplet of atom classes over all input files yields a decrease in the MAE by 0.88 kcal/mol to 3.20 kcal/mol and a ME of 17.31 kcal/mol. Estimating furthermore the empirical partial charge parameters  $q_i$  using Hirshfeld charges and ESP charges yields a MAE of 7.75 kcal/mol (ME of 17.89 kcal/mol) and a MAE of 4.52 kcal/mol (ME of 23.07 kcal/mol), respectively. Again, these energetic performances that are significantly worse must be considered unsatisfactory.

One interesting observation thereby concerns the calculated partial charges of the  $\text{Na}^+$  cation using the Hirshfeld partitioning scheme, as depicted in Figure 5.9: In contrast to the other atom types of the system that show a rather well-defined distribution of Hirshfeld charges, it appears that the Hirshfeld charge of the  $\text{Na}^+$  cation strongly depends on the number of oxygen atom ligands that are coordinated towards the cation, as illustrated in Figure 5.9(e). It is found that the calculated Hirshfeld charge  $q_{\text{Na}^+}$  of the sodium cation is in the range between 0.50 to 0.54 if three oxygen atoms are coordinated towards the metal cation. If two oxygen atoms are coordinated towards the  $\text{Na}^+$  cation,  $q_{\text{Na}^+}$  lies in the range between 0.58 to 0.63. If one oxygen atom and one or more additional hydrogen atoms are situated in the proximity of the  $\text{Na}^+$  cation proximity,  $q_{\text{Na}^+}$  lies in the range between 0.67 to 0.73. If only one oxygen atom is coordinated towards the  $\text{Na}^+$  cation and no hydrogen atoms are found in its proximity,  $q_{\text{Na}^+}$  lies in the range between 0.75 to 0.77. Interestingly, all estimated ESP partial charge values are found to be larger than these values, see Figure A.1 in Appendix A. These findings strongly indicate a varying atomic partial charge fluctuation depending on the number of oxygen atom ligands that is not taken into account using a fixed charge model as applied here. A similar effect is found concerning the oxygen atoms of the molecule, see Figure A.1 in Appendix A. In addition, possible changes in the covalent structure caused by the  $\text{Na}^+$  cation should be considered as well. An example of such an effect is illustrated in Figure 5.9(f): It is found that the distance between the oxygen and carbon atoms of the system varies depending on the number of oxygen atom ligands that are coordinated towards the sodium cation. If the oxygen atom is not coordinated towards the cation, the distribution is rather well-defined with an equilibrium distance around 1.23 Å. On the other hand, in case the oxygen atom is a ligand coordinated towards the  $\text{Na}^+$  cation, a small displacement is caused which results in C–O distances up to 1.26 Å. The simple bonds FF terms (see Equation (5.4)) that are of quadratic form are not able to take such effects into account. A refinement of the FF formulation itself in order to include such effects would be required. For such a task, the framework presented here might serve as a helpful tool in the future.

Using furthermore the empirical partial charge parameters of the original OPLS-AA FF and estimating the empirical vdW parameters  $\sigma_{ij}$  by applying the atomic Hirshfeld partitioning

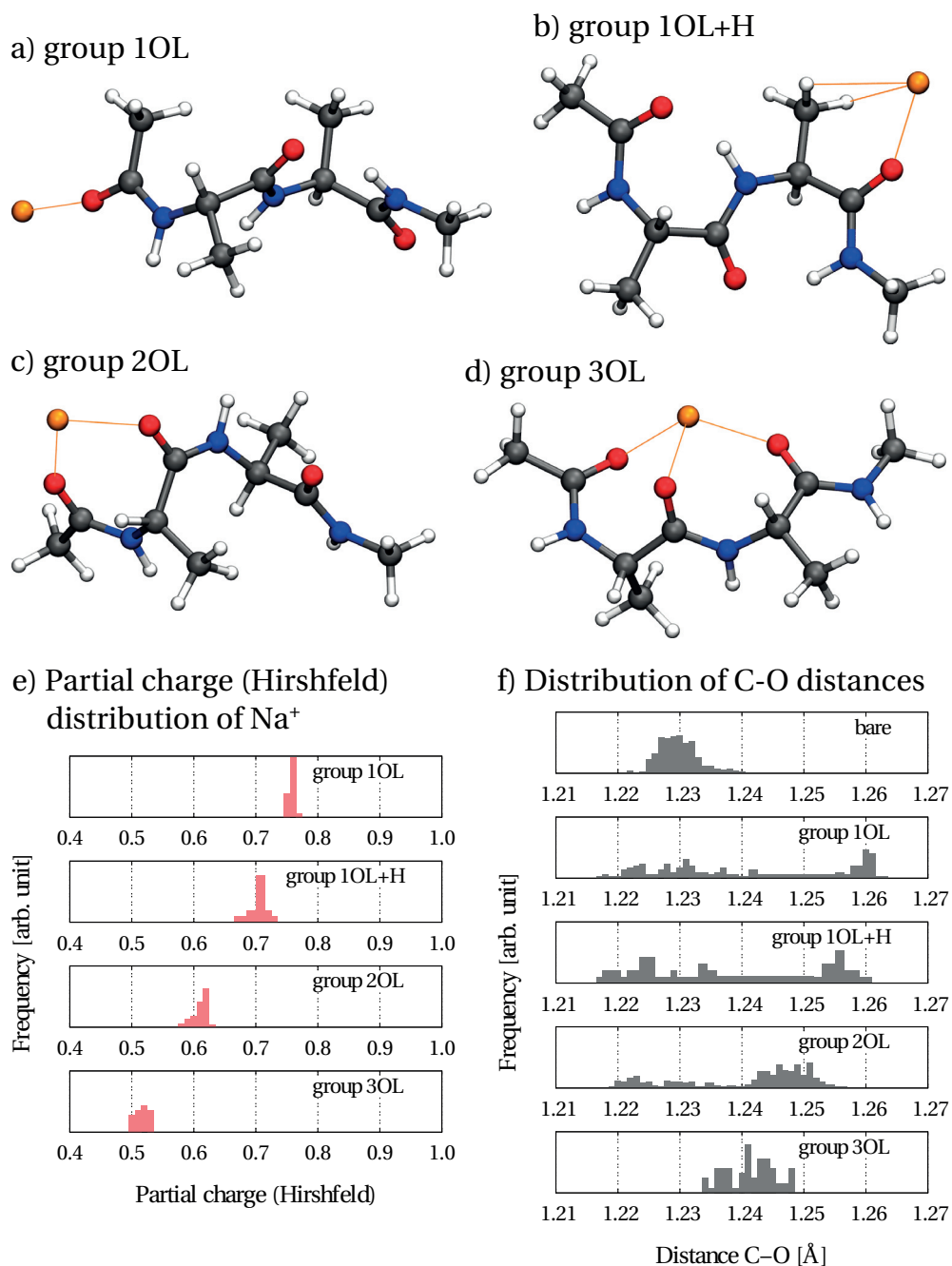


Figure 5.9 – Exemplary illustration of conformers of AcAla<sub>2</sub>NMe + Na<sup>+</sup> for which the sodium cation is surrounded by (a) one oxygen atom ligand (labeled conformer group 1OL), (b) one oxygen atom ligand as well as one or more hydrogen atoms in its proximity (labeled group 1OL+H), (c) two oxygen atom ligands (labeled group 2OL), and (d) three oxygen atom ligands (labeled group 3OL). (e) Distribution of the partial charges of the Na<sup>+</sup> cation calculated using the Hirshfeld partitioning scheme for the four different conformer groups. (f) Distribution of calculated distances between bonded oxygen and carbon atoms for the bare peptide and the four different conformer groups. The corresponding training set of conformers has been used for all cases.

Table 5.10 – Overview on the calculated MAEs and MEs in this section for both the systems AcAla<sub>2</sub>NMe and AcAla<sub>2</sub>NMe + Na<sup>+</sup> applying varying adjustment procedures and tackled FF parameters.

Tackled FF parameters and adjustment	AcAla <sub>2</sub> NMe		AcAla <sub>2</sub> NMe + Na <sup>+</sup>	
	MAE [kcal/mol]	ME [kcal/mol]	MAE [kcal/mol]	ME [kcal/mol]
• Standard OPLS-AA parameters	2.55	10.45	4.08	19.82
• $r_{ij}^0$ and $\theta_{ij}^0$ by averaging	1.81	8.93	3.20	17.31
• $r_{ij}^0$ and $\theta_{ij}^0$ by averaging • $q_i$ using Hirshfeld partitioning	2.62	10.02	7.75	17.89
• $r_{ij}^0$ and $\theta_{ij}^0$ by averaging • $q_i$ using ESP	2.93	10.17	4.52	23.07
• $r_{ij}^0$ and $\theta_{ij}^0$ by averaging • $\sigma_{ij}$ using atomic Hirshfeld partitioning • $\epsilon_{ij}$ using LASSO against $E^{\text{vdW}^{\text{TS}},\text{DFT}}$	1.69	5.17	2.80	16.52
• $r_{ij}^0$ and $\theta_{ij}^0$ by averaging • $\sigma_{ij}$ using atomic Hirshfeld partitioning • $\epsilon_{ij}$ using LASSO against $E_{\text{DFT}}^{\text{vdW}^{\text{TS}}}$ • $V_1^{ij}, V_2^{ij}, V_3^{ij}$ using LASSO against $E_{\text{DFT}}^{\text{tot}}$	0.76	2.81	1.87	15.14

scheme as well as estimating the  $\epsilon_{ij}$  parameters by using LASSO regression with a regularization parameter  $\lambda = 0.082$  and the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> energy  $E^{\text{vdW}^{\text{TS}},\text{DFT}}$  taken as response data yields a MAE of 2.80 kcal/mol and a ME of 16.52 kcal/mol. Finally, all torsions empirical parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  are estimated from the training set using LASSO regression with a regularization parameter  $\lambda = 0.018$  and the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data, which yields a MAE of 1.87 kcal/mol and a ME of 15.14 kcal/mol. As expected, the poor energetic performance of the FF due to the Coulomb interaction terms cannot be fully “recovered” by the adjustment of the torsional parameters alone, which results in an energetic description well above “chemical accuracy” when compared to the PBE+vdW<sup>TS</sup> DFT level.

Table 5.10 summarizes the MAEs and MEs calculated in this section for both the systems AcAla<sub>2</sub>NMe and AcAla<sub>2</sub>NMe + Na<sup>+</sup> applying the varying adjustment procedures and tackled FF parameters.

## 5.4 Conclusion and Outlook

A machine learning approach for modifying parameters of the standard OPLS-AA FF was proposed and laid out in detail. Besides using empirical partial charge parameters derived from the Hirshfeld partitioning scheme or reproduced from the electrostatic potential (ESP), the main focus of the procedure lies on deriving torsions or van der Waals parameters by simply fit-

ting the FF potential energy  $E_{\text{pot}}$  against high-level energies, *e.g.* from DFT calculations, using different regularized regression models. In particular, the LASSO regularization method shows promising results because it does not only serve as a form of shrinkage regression required to “detain” the FF parameters that serve as regression coefficients, but it also acts as a *de facto* variable selector by suppressing low-impact predictor variables. Applying the “Framework For Adjusting Force Fields Using Regularized Regression” (FFAFFURR)<sup>2</sup> and intending a *proof of principle* for the rather simple system of AcAla<sub>2</sub>NMe results in a MAE of 0.76 kcal/mol and the ME of 2.81 kcal/mol for the test set of minima conformers when compared to the PBE+vdW<sup>TS</sup> DFT level, which is a significant improvement when comparing to the MAE of 2.55 kcal/mol and the ME of 10.45 kcal/mol using the standard set of OPLS-AA FF parameters. Compared with the formulation of bonds and angles, torsions are the dominant degrees of freedom in the molecule. Hence, the main reason for the regression approach to be able to reproduce hierarchical energies of minima conformers within “chemical accuracy” must be found in the torsions terms of the potential energy formulation of the FF, as they provide sufficient flexibility in terms of the energetic description. Because parameters are derived using only energies at the DFT level as target data, the procedure allows for a fair assessment of how well the FF formulation itself is able to describe the potential energy. While for the rather simple system of AcAla<sub>2</sub>NMe it is possible to provide an energetic description within “chemical accuracy”, it appears that the general form of the conventional OPLS-AA FF used here is not suitable to provide a general energetic description for more challenging cases like peptide-cation systems. The reason for that is twofold: For one, the bonds and angles contributions only work rather well for minima on the PES. Describing arbitrary conformers that are not minima on the PES results in large discrepancies in the energetic description of the FF due to the quadratic form of the bonds and angles terms. For a more general description, one would need to adapt the procedure to include a more general form of the potential energy terms, *e.g.* the Morse potential form [407] being a candidate or additional terms of quartic or sextic form besides the standard quadratic form. Secondly, the Coulomb contributions must be considered problematic in terms of reproducing accurate conformational hierarchies, despite their physical nature. The reason for that is unclear although it has been found that standard DFT leads to errors in the electron density distribution in comparison to MP2 calculations for zwitterionic peptides [408]. Neither Hirshfeld estimates nor ESP derived parameters provide a reliable adjustment of the partial charge estimates, which is especially concerning when describing a more challenging system like AcAla<sub>2</sub>NMe + Na<sup>+</sup>. Adding a sodium cation to the system results in large energetic discrepancies with respect to energies calculated at the DFT level which cannot be “compensated” by the torsions terms that also do not contain parameters that are directly related to the added cation. The obtained MAE of 4.08 kcal/mol and the ME of 19.82 kcal/mol for the test set using standard OPLS-AA FF parameters when compared to the PBE+vdW<sup>TS</sup> level could only be reduced to 1.87 kcal/mol and 15.14 kcal/mol for the MAE and ME, respectively, when using adjusted FF parameters, meaning the energetic performance is still well above “chemical accuracy”. Suggestions for improvement include a cohesive study

---

<sup>2</sup>The code is available free of charge and can be downloaded from:  
<https://github.com/FHIBioGroup/ffaffurr-dev>

of how well different partial charge models, *e.g.* Hirshfeld, ESP, restricted (R)ESP [404], Mulliken [409], *etc.*, are able to reproduce conformational energy hierarchies for a wider range of systems if used as template charges in FFs. Of course, more sophisticated partial charge models like the Drude oscillator model [91,92] or fluctuating charge models [92,93,410] might be tackled in the future although the associated higher computational costs in comparison to simple fixed charge FF models need to be kept in mind.





## **6 Summary**

## Chapter 6. Summary

---

The goal of this thesis was to study gas-phase systems of bare peptides or in presence of metal cations. In the initial parts of this work, the focus was put on better understanding the “undamped” intramolecular interactions that shape peptides, thus shedding light on intrinsic structural motif propensities and bonding interactions. To that end, the peptide AcPheAla<sub>5</sub>LysH<sup>+</sup> was investigated in Chapter 3, a model system for studying helix formation in the gas phase, in order to fully understand the forces that stabilize the helical structure. In particular, the question was addressed of whether the local fixation of the positive charge at the peptide’s C-terminus is a prerequisite for forming helices by replacing the protonated C-terminal Lys residue by Ala and a sodium cation. The combination of gas-phase vibrational spectroscopy of cryogenically cooled ions with molecular simulations based on DFT allowed for detailed structure elucidation. It was found that the fixed location of the charge at the C-terminus is imperative for helix formation in peptides of this length in isolation, as this stabilizes the structure through a cation-helix dipole interaction. Interestingly, for sodiated AcPheAla<sub>6</sub> + Na<sup>+</sup> globular rather than helical structures were found caused by the strong cation-backbone and cation- $\pi$  interactions, leading to local distortions of peptide structure, preventing helix stabilization. A thorough comparison of experiment and theory revealed that even though the cation- $\pi$  interaction is energetically favored for AcPheAla<sub>6</sub> + Na<sup>+</sup> in the gas phase, the system remains kinetically trapped in a structural state that is characterized by cation-backbone interactions and that is energetically preferred in polar solvent.

The findings in Chapter 3 relied in part on the conformational search approach for finding the global minima of the gas-phase systems AcPheAla<sub>5</sub>LysH<sup>+</sup> and AcPheAla<sub>6</sub> + Na<sup>+</sup> that in turn relied on the usage of conventional force fields and different levels of DFA. Furthermore, the correct assignment of calculated IR spectra to their experimental counterparts was only possible when relying on computationally costly hybrid exchange-correlation functionals at the DFT level. This inspired a study presented in Chapter 4 where the goodness of commonly applied levels of theory, *i.e.* force fields (FFs), semi-empirical quantum chemistry methods, density-functional approximations (DFAs), composite methods, and wavefunction-based methods was being assessed and evaluated with respect to benchmark-grade coupled-cluster calculations. For the benchmark systems consisting of either a bare acetylhistidine or micro-solvated with a Zn<sup>2+</sup> cation it was found that force fields and semi-empirical methods are generally not reliable enough for an energetic description of these systems within “chemical accuracy” of 1 kcal/mol. While the energetic performance of GGA xc functionals like PBE and BLYP, as well as the composite method PBEh-3c, was above “chemical accuracy” for systems containing a Zn<sup>2+</sup> cation, it was found that hybrid xc functionals performed generally well with small energetic deviations within 1 kcal/mol. Out of all tested methods, the double hybrid xc functional B3LYP+XYG3 and the wavefunction-based MP2 method resembled the benchmark method DLPNO-CCSD(T) best.

Taking the findings of Chapter 4 into account, in particular the poor energetic performance of FFs, a necessity to be able to adjust parameters of a particular FF for a specific system in question using minimal effort was realized. A framework for a machine learning approach was introduced in Chapter 5 that aims to modify the FF parameters in such a way that energy

---

hierarchies obtained using DFT (or any other high-level method) are to be reproduced within “chemical accuracy” as well as being rather simplistic in order to be able to be undergone by the end-users themselves. The “Framework For Adjusting Force Fields Using Regularized Regression” (FFAFFURR) is able to modify van der Waals parameters and torsional parameters in the potential-energy function  $E_{\text{pot}}$  of a particular FF, here OPLS-AA, by fitting  $E_{\text{pot}}$  against high-level energies, *e.g.* from DFT calculations, using different regularized regression models for a rather small subset out of a large pool of conformers. In particular, the LASSO regularization method showed promising results because it not only serves as a form of shrinkage regression required to “detain” the FF parameters that serve as regression coefficients, but it also acts as a *de facto* variable selector by suppressing low-impact predictor variables. Furthermore, partial charge parameters can be derived from the Hirshfeld partitioning scheme or reproduced from the electrostatic potential (ESP). A *proof of principle* for the system of AcAla<sub>2</sub>NMe resulted in an energetic description that yields mean deviations within “chemical accuracy” when compared to the PBE+vdW<sup>TS</sup> DFT level, which is a significant improvement when comparing to the rather poor energetic description provided when using the standard set of OPLS-AA FF parameters. Because parameters were derived using only energies at the DFT level as target data, the procedure allows for a fair assessment of how well the FF formulation itself is able to describe the potential energy. For the rather simple system of AcAla<sub>2</sub>NMe, the torsions terms of the FF formulation are able to provide sufficient “flexibility” in terms of the energetic description of the molecule in order to “compensate” shortcomings in the energetic description caused by the other terms of the potential energy function. In general however, it appears that the form of the conventional OPLS-AA FF is not suitable to provide an accurate enough energetic description for a more challenging system. For one, bonds and angles contributions of quadratic form are only suited to energetically describe minima on the PES. Secondly, the Coulomb contributions must be considered problematic in terms of reproducing accurate conformational hierarchies, despite their physical nature. The reason for that is unclear and requires further investigation. Neither Hirshfeld estimates nor ESP derived parameters provide a reliable adjustment of the partial charge estimates, which is especially concerning when describing a peptide-cation system like AcAla<sub>2</sub>NMe + Na<sup>+</sup> for which Coulomb contributions play an even more important role. With the procedure in itself working as shown with the *proof of principle* for AcAla<sub>2</sub>NMe, it may serve as a stepping stone for further improving the formulation of FF potential energy functions in order to yield a more accurate energetic description for such systems.



# A Appendix: Listing of Force Field Parameters for AcAla<sub>2</sub>NMe + Na<sup>+</sup>

In Chapter 5.3, the study of “adjusting” empirical parameters of the OPLS-AA FF for the system of AcAla<sub>2</sub>NMe has been repeated for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup>. In the following, a detailed listing of all values of “adjusted” empirical parameters as well as plotted regression coefficients, the *RSS*, and calculated MAEs and MEs for the test set obtained by using different regression models is provided.

Table A.1 – Original OPLS-AA FF parameters  $r_{ij}^0$  and  $\theta_{ij}^0$  and their “adjusted” counterparts for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup> obtained by averaging over all corresponding atomic pairwise distances and angles of the same pair or triplet of atom classes over all input files.

Atom class pair	$r_{ij}^0$ [Å] (original FF)	$r_{ij}^0$ [Å] (“adjusted” FF)	Atom class triplet	$\theta_{ij}^0$ [°] (original FF)	$\theta_{ij}^0$ [°] (“adjusted” FF)
(3, 4)	1.2290	1.2403	(4, 3, 13)	120.4000	120.5641
(3, 13)	1.5220	1.5353	(4, 3, 24)	122.9000	121.2738
(3, 24)	1.3350	1.3578	(13, 3, 24)	116.6000	118.1161
(13, 13)	1.5290	1.5320	(3, 13, 13)	111.1000	112.4653
(13, 24)	1.4490	1.4619	(3, 13, 24)	110.1000	110.1438
(13, 46)	1.0900	1.0970	(13, 13, 24)	109.7000	112.3959
(24, 45)	1.0100	1.0187	(3, 13, 46)	109.5000	108.7366
			(13, 13, 46)	110.7000	110.1281
			(24, 13, 46)	109.5000	108.8104
			(46, 13, 46)	107.8000	108.3819
			(3, 24, 13)	121.9000	127.1322
			(3, 24, 45)	119.8000	114.6338
			(13, 24, 45)	118.4000	116.0534

## Appendix A. Appendix: Listing of Force Field Parameters for AcAla<sub>2</sub>NMe + Na<sup>+</sup>

Table A.2 – Original OPLS-AA FF parameters  $q_i$  and their “adjusted” counterparts for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup> obtained by averaging over all Hirshfeld or ESP charges of the same atom type over all input files.

Atom type	$q_i$ (original FF)	$q_i$ (Hirshfeld) (“adjusted” FF)	$q_i$ (ESP) (“adjusted” FF)
80	-0.1800	-0.1048	-0.5285
85	0.0600	0.0561	0.1458
166	0.1400	0.0223	0.3267
177	0.5000	0.1635	0.4622
178	-0.5000	-0.2596	-0.5634
180	-0.5000	-0.0785	-0.3893
183	0.3000	0.1324	0.2838
184	0.0200	-0.0438	-0.4038
349	1.0000	0.6551	0.9149

Table A.3 – Original OPLS-AA FF parameters  $\sigma_{ij}$  and their “adjusted” counterparts for the system of AcAla<sub>2</sub>NMe+Na<sup>+</sup> obtained by the atomic Hirshfeld partitioning scheme as explained in Subsection 5.2.3.

Atom type pair	$\sigma_{ij}$ [Å] (original FF)	$\sigma_{ij}$ [Å] (“adjusted” FF)	Atom type pair	$\sigma_{ij}$ [Å] (original FF)	$\sigma_{ij}$ [Å] (“adjusted” FF)
(80, 80)	3.5000	2.9249	(166, 178)	3.2187	2.8847
(80, 85)	2.9580	2.6317	(166, 180)	3.3727	2.9013
(80, 166)	3.5000	2.9615	(166, 184)	3.5000	2.9462
(80, 177)	3.6228	2.9647	(166, 349)	3.7743	2.6765
(80, 178)	3.2187	2.8481	(177, 177)	3.7500	3.0045
(80, 180)	3.3727	2.8647	(177, 178)	3.3317	2.8879
(80, 184)	3.5000	2.9096	(177, 180)	3.4911	2.9044
(80, 349)	3.7743	2.6399	(177, 184)	3.6228	2.9493
(85, 85)	2.5000	2.3385	(177, 349)	3.9067	2.6797
(85, 166)	2.9580	2.6683	(178, 178)	2.9600	2.7713
(85, 177)	3.0619	2.6715	(178, 180)	3.1016	2.7879
(85, 178)	2.7203	2.5549	(178, 184)	3.2187	2.8328
(85, 180)	2.8504	2.5715	(178, 349)	3.4709	2.5631
(85, 184)	2.9580	2.6163	(180, 180)	3.2500	2.8044
(85, 349)	3.1898	2.3467	(180, 184)	3.3727	2.8493
(166, 166)	3.5000	2.9982	(180, 349)	3.6370	2.5797
(166, 177)	3.6228	3.0013	(184, 349)	3.7743	2.6245

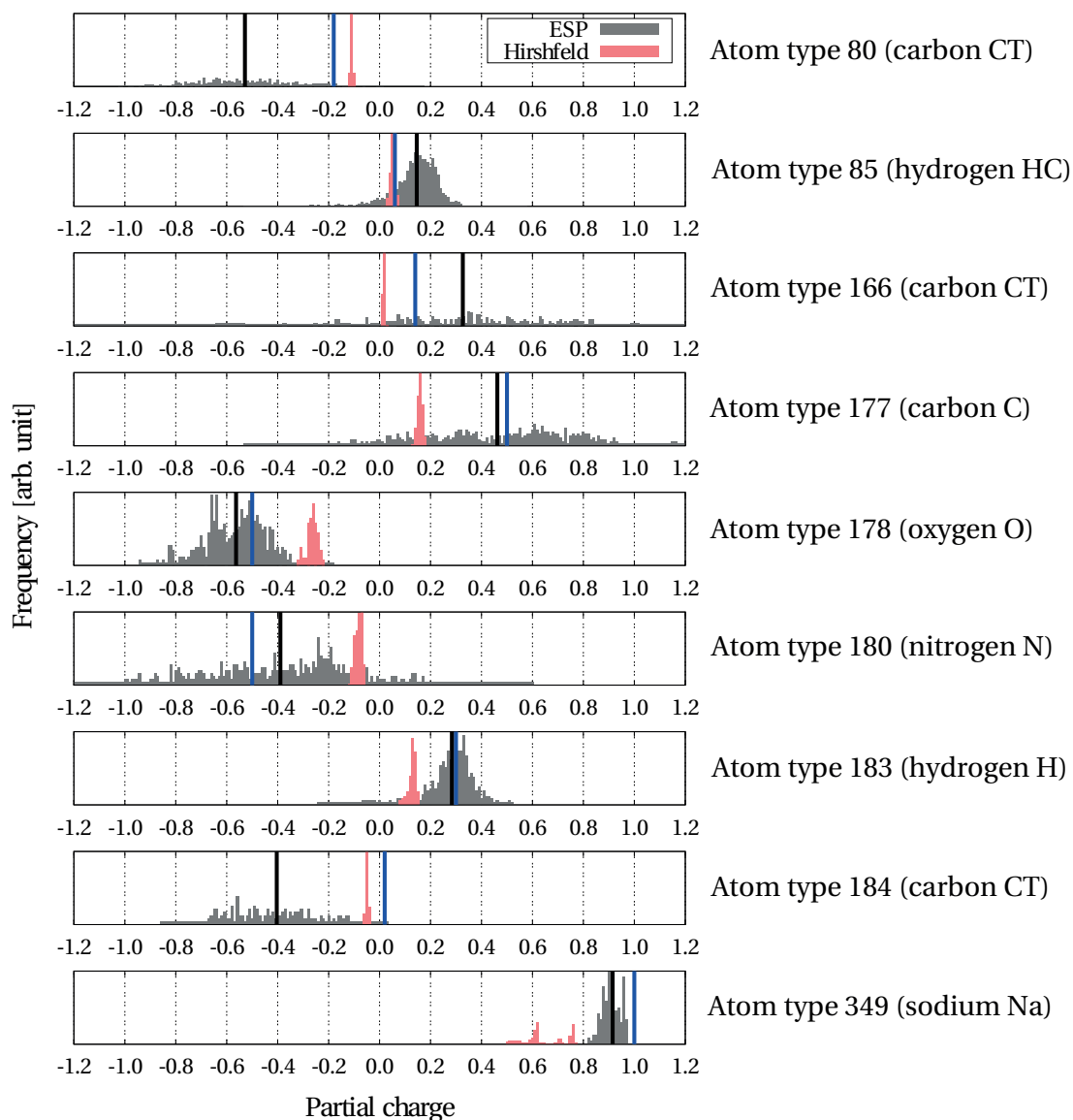


Figure A.1 – Distributions of Hirshfeld (red) and ESP (gray) charges over the training set of conformers for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup>. For visibility purposes, the frequencies of the ESP charges have been scaled by a factor of 3. Vertical blue lines denote partial charge parameters of the original OPLS-AA FF while vertical black lines denote the average of the ESP values.

## Appendix A. Appendix: Listing of Force Field Parameters for AcAla<sub>2</sub>NMe + Na<sup>+</sup>

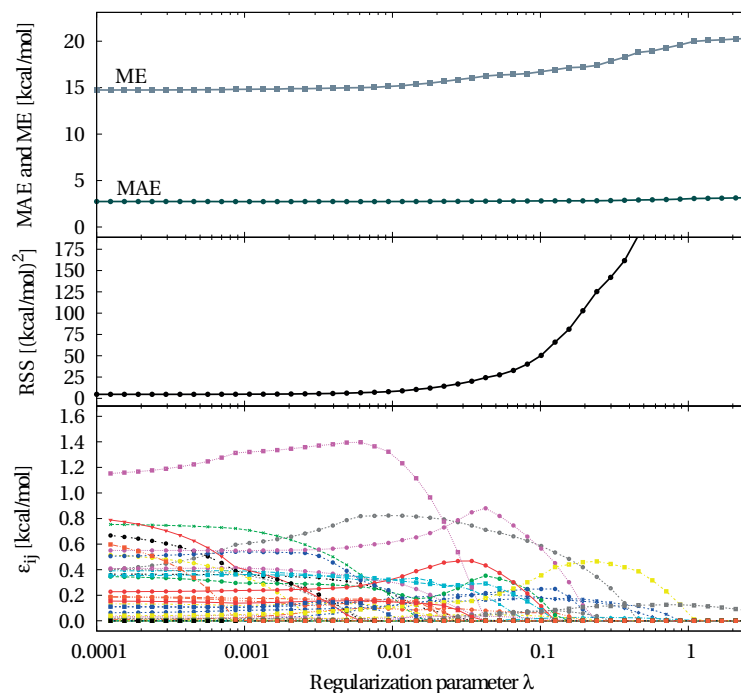


Figure A.2 – Estimated regression coefficients  $\epsilon_{ij}$  (bottom), the residual sum of squares  $RSS$  (middle), and calculated MAEs and MEs (top) for the test set of AcAla<sub>2</sub>NMe + Na<sup>+</sup> obtained by using LASSO regression with the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> energy  $E^{vdW^{TS},DFT}$  taken as response data.

Table A.4 – Original OPLS-AA FF parameters  $\epsilon_{ij}$  and their “adjusted” counterparts for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup> obtained by using LASSO regression ( $\lambda = 0.082$ ) with the pairwise Tkatchenko-Scheffler vdW<sup>TS</sup> energy  $E^{vdW^{TS},DFT}$  taken as response data.

Atom type pair	$\epsilon_{ij}$ [kcal/mol] (original FF)	$\epsilon_{ij}$ [kcal/mol] (“adjusted” FF)	Atom type pair	$\epsilon_{ij}$ [kcal/mol] (original FF)	$\epsilon_{ij}$ [kcal/mol] (“adjusted” FF)
(80, 80)	0.0660	0.0000	(166, 178)	0.1177	0.2473
(80, 85)	0.0445	0.1769	(166, 180)	0.1059	0.0000
(80, 166)	0.0660	0.0000	(166, 184)	0.0660	0.0000
(80, 177)	0.0832	0.0000	(166, 349)	0.0057	0.0000
(80, 178)	0.1177	0.0000	(177, 177)	0.1050	0.0000
(80, 180)	0.1059	0.0000	(177, 178)	0.1485	0.0745
(80, 184)	0.0660	0.0000	(177, 180)	0.1336	0.2140
(80, 349)	0.0057	0.0000	(177, 184)	0.0832	0.0000
(85, 85)	0.0300	0.0000	(177, 349)	0.0072	0.6587
(85, 166)	0.0445	0.1920	(178, 178)	0.2100	0.1515
(85, 177)	0.0561	0.0240	(178, 180)	0.1889	0.3012
(85, 178)	0.0794	0.0138	(178, 184)	0.1177	0.0000
(85, 180)	0.0714	0.0420	(178, 349)	0.0102	0.0656
(85, 184)	0.0445	0.0629	(180, 180)	0.1700	0.0000
(85, 349)	0.0039	0.6262	(180, 184)	0.1059	0.0000
(166, 166)	0.0660	0.0000	(180, 349)	0.0092	0.0000
(166, 177)	0.0832	0.2485	(184, 349)	0.0057	0.0000



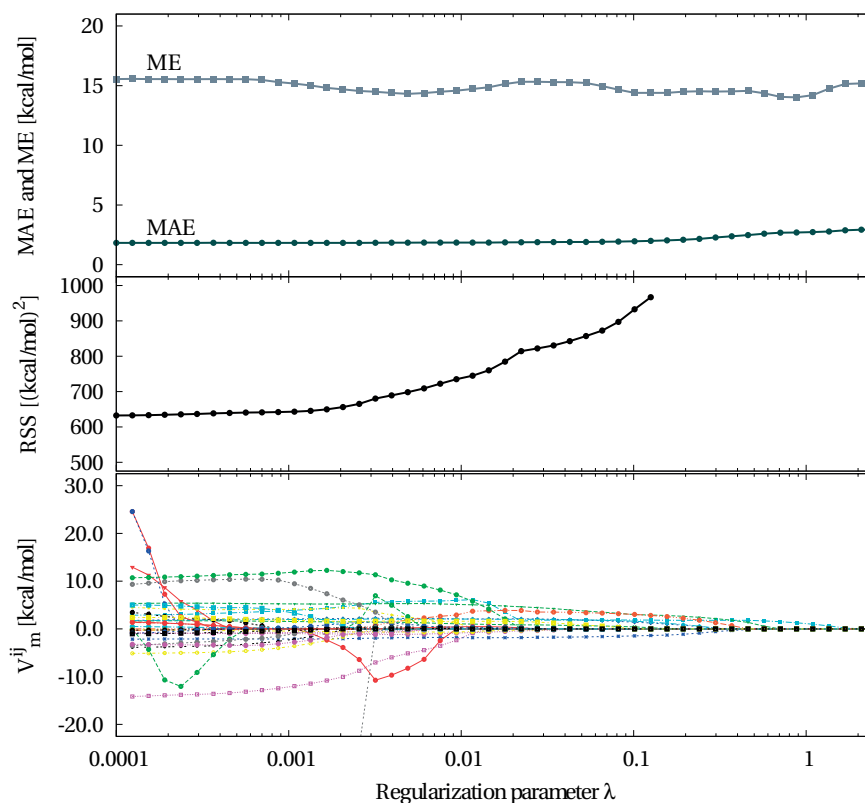


Figure A.3 – Estimated regression coefficients  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  (bottom), the residual sum of squares  $RSS$  (middle), and calculated MAEs and MEs (top) for the test set of AcAla<sub>2</sub>NMe + Na<sup>+</sup> obtained by using LASSO regression with the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data.

Table A.5 – Original OPLS-AA FF parameters  $V_1^{ij}$ ,  $V_2^{ij}$ , and  $V_3^{ij}$  as well as their “adjusted” counterparts for the system of AcAla<sub>2</sub>NMe + Na<sup>+</sup> obtained by using LASSO regression ( $\lambda = 0.018$ ) with the total DFT energy of the PBE+vdW<sup>TS</sup> model taken as response data.

Atom class quadruplett	$V_1^{ij}$ [kcal/mol] (original FF)	$V_1^{ij}$ [kcal/mol] (“adjusted” FF)	$V_2^{ij}$ [kcal/mol] (original FF)	$V_2^{ij}$ [kcal/mol] (“adjusted” FF)	$V_3^{ij}$ [kcal/mol] (original FF)	$V_3^{ij}$ [kcal/mol] (“adjusted” FF)
(24,3,13,13)	1.173	0.884	0.189	0.000	-1.200	-0.285
(24,3,13,24)	1.816	-1.838	1.222	2.018	1.581	0.000
(4,3,24,13)	0.000	0.000	6.089	0.000	0.000	0.000
(4,3,24,45)	0.000	3.485	4.900	0.000	0.000	0.000
(13,3,24,13)	2.300	0.000	6.089	3.867	0.000	0.375
(13,3,24,45)	0.000	0.000	4.900	0.000	0.000	2.083
(3,13,13,46)	0.000	0.000	0.000	0.000	-0.100	0.261
(24,13,13,46)	0.000	0.000	0.000	0.000	0.464	0.000
(46,13,13,46)	0.000	0.000	0.000	0.000	0.300	0.000
(3,13,24,3)	-2.365	4.787	0.912	-0.533	-0.850	1.579
(13,13,24,3)	0.000	2.102	0.462	0.000	0.000	0.000



# Bibliography

- [1] I. Bertini, H. B. Gray, E. I. Stiefel, and J. S. Valentine. *Biological Inorganic Chemistry*. University Science Books, 2007.
- [2] K. J. Waldron, J. C. Rutherford, D. Ford, and N. J. Robinson. Metalloproteins and Metal Sensing. *Nature*, 460(7257):823–30, 2009.
- [3] K. P. Kepp. Bioinorganic Chemistry of Alzheimer's Disease. *Chem. Rev.*, 112(10):5193–5239, 2012.
- [4] S. Zirah, S. A. Kozin, A. K. Mazur, A. Blond, M. Cheminant, I. Ségalas-Milazzo, P. Debey, and S. Rebuffat. Structural Changes of Region 1-16 of the Alzheimer Disease Amyloid  $\beta$ -Peptide upon Zinc Binding and in Vitro Aging. *J. Biol. Chem.*, 281(4):2151–2161, 2006.
- [5] H. Sandstead. Understanding Zinc: Recent Observations and Interpretations. *J. Lab. Clin. Med.*, 124(3):322–327, 1994.
- [6] S. Lindskog. Structure and Mechanism of Carbonic Anhydrase. *Pharmacol. Therapeut.*, 74(1):1–20, 1997.
- [7] A. E. Eriksson, T. A. Jones, and A. Liljas. Refined Structure of Human Carbonic Anhydrase II at 2.0 Å Resolution. *Proteins: Struct., Funct., Bioinf.*, 4(4):274–282, 1988.
- [8] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996.
- [9] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136:B864–B871, 1964.
- [10] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140:A1133–A1138, 1965.
- [11] O. Guvench and A. D. MacKerell. *Comparison of Protein Force Fields for Molecular Dynamics Simulations*, In A. Kukol (editor), *Molecular Modeling of Proteins*, pages 63–88. Humana Press, 2008.
- [12] W. L. Jorgensen and J. Tirado-Rives. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

## Bibliography

---

- [13] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.
- [14] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B*, 105(28):6474–6487, 2001.
- [15] A. N. Tikhonov, V. Y. Arsenin, and F. John. *Solutions of Ill-Posed Problems*, Volume 14. Winston, 1977.
- [16] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola. *Numerical Methods for the Solution of Ill-Posed Problems*, Volume 328. Springer Science & Business Media, 2013.
- [17] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.
- [18] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. Roy. Stat. Soc. B Met.*, 58(1):267–288, 1996.
- [19] A. Galuschka. *Biochemie für Ahnungslose – Eine Einstiegshilfe für Studierende*. Hirzel Verlag, 2015.
- [20] T. E. Creighton. *Proteins: Structures and Molecular Properties*. Macmillan, 1993.
- [21] P. D. Bailey. *An Introduction to Peptide Chemistry*. Wiley, Salle Sauerländer, 1990.
- [22] H.-D. Jakubke and H. Jeschkeit. *Aminosäuren, Peptide, Proteine: Eine Einführung*. Akademie-Verlag, 1973.
- [23] A. Guijarro and M. Yus. *The Origin of Chirality in the Molecules of Life: A Revision From Awareness to the Current Theories and Perspectives of This Unsolved Problem*. Royal Society of Chemistry, 2008.
- [24] F. W. J. Teale and G. Weber. Ultraviolet Fluorescence of the Aromatic Amino Acids. *Biochem. J.*, 65(3):476–482, 1957.
- [25] D. B. Wetlaufer. Ultraviolet Spectra of Proteins and Amino Acids. Volume 17 of *Advances in Protein Chemistry*, pages 303–390. Academic Press, 1963.
- [26] M. Tanokura. <sup>1</sup>H-NMR Study on the Tautomerism of the Imidazole Ring of Histidine Residues: I. Microscopic pK Values and Molar Ratios of Tautomers in Histidine-Containing Peptides. *BBA-Protein Struct. M.*, 742(3):576–585, 1983.
- [27] E. A. Barnard and W. D. Stein. The Roles of Imidazole in Biological Systems. *Adv. Enzymol. Rel. S. Bi.*, 20:51–110, 1958.

- [28] W. Müller-Esterl. *Biochemie. Eine Einführung für Mediziner und Naturwissenschaftler*. Springer Spektrum, 2004.
- [29] N. Sewald and H. D. Jakubke. *Peptides: Chemistry and Biology*. Wiley, 2015.
- [30] G. N. Ramachandran and V. Sasisekharan. Conformation of Polypeptides and Proteins. Volume 23 of *Advances in Protein Chemistry*, pages 283–437. Academic Press, 1968.
- [31] I. Langmuir. The Arrangement of Electrons in Atoms and Molecules. *J. Am. Chem. Soc.*, 41(6):868–934, 1919.
- [32] G. N. Lewis. The Atom and the Molecule. *J. Am. Chem. Soc.*, 38(4):762–785, 1916.
- [33] M. Williamson. *How Proteins Work*. Garland Science, 2011.
- [34] P. A. Kollman. Noncovalent Interactions. *Acc. Chem. Res.*, 10(10):365–371, 1977.
- [35] P. Hobza, R. Zahradník, and K. Müller-Dethlefs. The World of Non-Covalent Interactions: 2006. *Collect. Czech. Chem. C.*, 71(4):443–531, 2006.
- [36] W. Pauli. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Z. Phys.*, 31(1):765–783, 1925.
- [37] L. Pauling. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*, Volume 18. Cornell University Press, 1960.
- [38] S. S. Batsanov. Van der Waals Radii of Elements. *Inorg. Mater.*, 37(9):871–885, 2001.
- [39] A. Bondi. Van der Waals Volumes and Radii. *J. Phys. Chem.*, 68(3):441–451, 1964.
- [40] L. Pauling. The Nature of the Chemical Bond. Application of Results Obtained from the Quantum Mechanics and from a Theory of Paramagnetic Susceptibility to the Structure of Molecules. *J. Am. Chem. Soc.*, 53(4):1367–1400, 1931.
- [41] Dougherty D. A. The Cation- $\pi$  Interaction. *Acc. Chem. Res.*, 46(4):885–893, 2013.
- [42] Z. Shi, C. A. Olson, and N. R. Kallenbach. Cation- $\pi$  Interaction in Model  $\alpha$ -Helical Peptides. *J. Am. Chem. Soc.*, 124(13):3284–3291, 2002.
- [43] F. London. Zur Theorie und Systematik der Molekularkräfte. *Z. Phys.*, 63(3):245–279, 1930.
- [44] R. Eisenschitz and F. London. Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften. *Z. Phys.*, 60(7):491–527, 1930.
- [45] J. E. Jones. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *P. Roy. Soc. Lond. A Mat.*, 106(738):463–477, 1924.

## Bibliography

---

- [46] J. Kuriyan, B. Konforti, and D. Wemmer. *The Molecules of Life: Physical and Chemical Principles*. Garland Science, 2012.
- [47] G. C. Pimentel and A. L. McClellan. *The Hydrogen Bond*. A Series of Chemistry Books. W. H. Freeman, 1960.
- [48] K. Morokuma. Why Do Molecules Interact? The Origin of Electron Donor-Acceptor Complexes, Hydrogen Bonding and Proton Affinity. *Acc. Chem. Res.*, 10(8):294–300, 1977.
- [49] L. G. Vanquickenborne. *Quantum Chemistry of the Hydrogen Bond*, In P. L. Huyskens, W. A. P. Luck, and T. Zeegers-Huyskens (editors), *Intermolecular Forces: An Introduction to Modern Methods and Results*, pages 31–53. Springer, 1991.
- [50] G. A. Jeffrey and W. Saenger. *Hydrogen Bonding in Biological Structures*. Springer Science & Business Media, 2012.
- [51] J. J. Dannenberg. Cooperativity in Hydrogen Bonded Aggregates. Models for Crystals and Peptides. *J. Mol. Struct.*, 615(1):219–226, 2002.
- [52] K. U. Linderstrøm-Lang. *Lane Medical Lectures: Proteins and Enzymes*. Med. Sciences: Stanford University Publications / Univ. Series. Stanford University Press, 1952.
- [53] J. Rossjohn, R. Cappai, S. C. Feil, A. Henry, W. J. McKinstry, D. Galatis, L. Hesse, G. Multhaup, K. Beyreuther, C. L. Masters, and M. W. Parker. Crystal Structure of the N-Terminal, Growth Factor-Like Domain of Alzheimer Amyloid Precursor Protein. *Nat. Struct. Mol. Biol.*, 6(4):327–331, 1999.
- [54] C. N. Pace and J. M. Scholtz. A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophys J.*, 75(1):422–427, 1998.
- [55] L. Pauling, R. B. Corey, and H. R. Branson. The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *P. Natl. Acad. Sci. USA*, 37(4):205–211, 1951.
- [56] M. Crisma, F. Formaggio, A. Moretto, and C. Toniolo. Peptide Helices Based on  $\alpha$ -Amino Acids. *Biopolymers*, 84(1):3–12, 2006.
- [57] R. B. Cooley, D. J. Arp, and P. A. Karplus. Evolutionary Origin of a Secondary Structure:  $\pi$ -Helices as Cryptic but Widespread Insertional Variations of  $\alpha$ -Helices That Enhance Protein Functionality. *J. Mol. Biol.*, 404(2):232–246, 2010.
- [58] A. A. Adzhubei and M. J. E. Sternberg. Left-Handed Polyproline II Helices Commonly Occur in Globular Proteins. *J. Mol. Biol.*, 229(2):472–493, 1993.
- [59] F. S. Nandel and R. Jaswal. New Type of Helix and  $2_7$  Ribbon Structure Formation in Poly  $\Delta$ Leu Peptides: Construction of a Single-Handed Template. *Biomacromolecules*, 8(10):3093–3101, 2007.

- [60] L. Pauling and R. B. Corey. The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *Proc. Natl. Acad. Sci. USA*, 37(5):251–256, 1951.
- [61] V. Pavone, G. Gaeta, A. Lombardi, F. Natri, O. Maglio, C. Isernia, and M. Saviano. Discovering Protein Secondary Structures: Classification and Description of Isolated  $\alpha$ -Turns. *Biopolymers*, 38(6):705–721, 1996.
- [62] C. M. Venkatachalam. Stereochemical Criteria for Polypeptides and Proteins. V. Conformation of a System of Three Linked Peptide Units. *Biopolymers*, 6(10):1425–1436, 1968.
- [63] K. Möhle, M. Gußmann, and H.-J. Hofmann. Structural and Energetic Relations Between  $\beta$  turns. *J. Comput. Chem.*, 18(11):1415–1430, 1997.
- [64] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. Stereochemical Quality of Protein Structure Coordinates. *Proteins*, 12(4):345–364, 1992.
- [65] J. N. Onuchic and P. G. Wolynes. Theory of Protein Folding. *Curr. Opin. Struc. Biol.*, 14(1):70–75, 2004.
- [66] K. A. Dill and J. L. MacCallum. The Protein-Folding Problem, 50 Years On. *Science*, 338(6110):1042–1046, 2012.
- [67] M. Karplus. The Levinthal Paradox: Yesterday and Today. *Fold. Des.*, 2(1):S69–S75, 1997.
- [68] C. Levinthal. *How to Fold Graciously*. Debrunner P., Tsibris J. C. M., Münck E. (Eds.), Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House, Monticello, Illinois. University of Illinois Press, Urbana, 1969.
- [69] D. Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge Molecular Science. Cambridge University Press, 2003.
- [70] C. Levinthal. Are There Pathways for Protein Folding? *J. Chim. Phys.*, 65:44–45, 1968.
- [71] K. A. Dill and H. S. Chan. From Levinthal to Pathways to Funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.
- [72] C. B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973.
- [73] S. Govindarajan and R. A. Goldstein. On the Thermodynamic Hypothesis of Protein Folding. *Proc. Natl. Acad. Sci. USA*, 95(10):5545–5549, 1998.
- [74] N. D. Socci, J. N. Onuchic, and P. G. Wolynes. Diffusive Dynamics of the Reaction Coordinate for Protein Folding Funnels. *J. Chem. Phys.*, 104(15):5860–5868, 1996.
- [75] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.*, 48(1):545–600, 1997.

## Bibliography

---

- [76] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein Folding Funnels: A Kinetic Approach to the Sequence-Structure Relationship. *Proc. Natl. Acad. Sci. USA*, 89(18):8721–8725, 1992.
- [77] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Struct., Funct., Bioinf.*, 21(3):167–195, 1995.
- [78] D. J. Brockwell and S. E. Radford. Intermediates: Ubiquitous Species on Folding Energy Landscapes? *Curr. Opin. Struct. Biol.*, 17(1):30–37, 2007.
- [79] C. L. Brooks, J. N. Onuchic, and D. J. Wales. Taking a Walk on a Landscape. *Science*, 293(5530):612–613, 2001.
- [80] H. von Helmholtz, J. W. Hittorf, and J. D. Waals. *Physical Memoirs Selected and Translated from Foreign Sources*. Taylor & Francis, 1888.
- [81] I. M. Klotz and R. M. Rosenberg. *Chemical Thermodynamics: Basic Concepts and Methods*. Wiley, 2008.
- [82] J. W. Gibbs. A Method of Geometrical Representation of the Thermodynamic Properties of Substances by Means of Surfaces. *Trans. Conn. Acad.*, 2:382–404, 1873.
- [83] T. Heimburg. *Thermal Biophysics of Membranes*. Tutorials in Biophysics. Wiley, 2008.
- [84] I. N. Levine. *Quantum Chemistry*. Prentice Hall, 2009.
- [85] R. G. Parr. On the genesis of a theory. *Int. J. Quantum Chem.*, 37(4):327–347, 1990.
- [86] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.
- [87] J. Wang, P. Cieplak, and P. A. Kollman. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.*, 21(12):1049–1074, 2000.
- [88] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.
- [89] G. Kaminski, E. M. Duffy, T. Matsui, and W. L. Jorgensen. Free Energies of Hydration and Pure Liquid Properties of Hydrocarbons from the OPLS All-Atom Model. *J. Phys. Chem.*, 98(49):13077–13082, 1994.



- [90] H. S. Antila and E. Salonen. *Polarizable Force Fields*, In L. Monticelli and E. Salonen (editors), *Biomolecular Simulations: Methods and Protocols*, pages 215–241. Humana Press, Totowa, NJ, 2013.
- [91] P. Drude, C. Riborg, and R. A. Millikan. *The Theory of Optics... Translated from German by C. R. Mann and R. A. Millikan*. Longmans, Green & Company, 1902.
- [92] P. E. M. Lopes, B. Roux, and A. D. MacKerell. Molecular Modeling and Dynamics Studies with Explicit Inclusion of Electronic Polarizability: Theory and Applications. *Theor. Chem. Acc.*, 124(1):11–28, 2009.
- [93] A. K. Rappe and W. A. Goddard. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.*, 95(8):3358–3363, 1991.
- [94] P. Ren and J. W. Ponder. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B*, 107(24):5933–5947, 2003.
- [95] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr., M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B*, 114(8):2549–2564, 2010.
- [96] P. Ren, C. Wu, and J. W. Ponder. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.*, 7(10):3143–3161, 2011.
- [97] N. L. Allinger, Y. H. Yuh, and J. H. Lii. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. *J. Am. Chem. Soc.*, 111(23):8551–8566, 1989.
- [98] T. A. Halgren. The Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and vdW Parameters. *J. Am. Chem. Soc.*, 114(20):7827–7843, 1992.
- [99] J. C. Wu, J.-P. Piquemal, R. Chaudret, P. Reinhardt, and P. Ren. Polarizable Molecular Dynamics Simulation of Zn(II) in Water Using the AMOEBA Force Field. *J. Chem. Theory Comput.*, 6(7):2059–2070, 2010.
- [100] Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.*, 9(9):4046–4063, 2013.
- [101] E. Penev, J. Ireta, and J.-E. Shea. Energetics of Infinite Homopolypeptide Chains: A New Look at Commonly Used Force Fields. *J. Phys. Chem. B*, 112(22):6872–6877, 2008.
- [102] M. Kolář, K. Berka, P. Jurečka, and P. Hobza. On the Reliability of the AMBER Force Field and its Empirical Dispersion Contribution for the Description of Noncovalent Complexes. *ChemPhysChem*, 11(11):2399–2408, 2010.

## Bibliography

---

- [103] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic Validation of Protein Force Fields against Experimental Data. *PLOS ONE*, 7(2):1–6, 2012.
- [104] G. Collier, N. A. Vellore, J. A. Yancey, S. J. Stuart, and R. A. Latour. Comparison Between Empirical Protein Force Fields for the Simulation of the Adsorption Behavior of Structured LK Peptides on Functionalized Surfaces. *Biointerphases*, 7(1):24, 2012.
- [105] E. Schrödinger. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.*, 28:1049–1070, 1926.
- [106] M. Springborg. *Methods of Electronic-Structure Calculations*. Wiley, 2000.
- [107] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Ann. Phys.*, 389(20):457–484, 1927.
- [108] J. Goodisman. *Diatomic Interaction Potential Theory*. Elsevier Science, 1973.
- [109] B. T. Sutcliffe. The Coupling of Nuclear and Electronic Motions in Molecules. *J. Chem. Soc. Faraday Trans.*, 89:2321–2335, 1993.
- [110] P. J. Mohr, D. B. Newell, and B. N. Taylor. CODATA Recommended Values of the Fundamental Physical Constants: 2014. *Rev. Mod. Phys.*, 88:035009, 2016.
- [111] J. O. Hirschfelder and W. J. Meath. The Nature of Intermolecular Forces. In *Advances in Chemical Physics*, Volume 12, pages 3–106. Wiley, 1967.
- [112] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *P. Camb. Philos. Soc.*, 24(1):111–132, 1928.
- [113] R. McWeeny. Natural Units in Atomic and Molecular Physics. *Nature*, 243:196–198, 1973.
- [114] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. International Series of Monographs on Chemistry. Oxford University Press, 1994.
- [115] G. B. Arfken. *Mathematical Methods for Physicists*. Elsevier Science, 2013.
- [116] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *P. Camb. Philos. Soc.*, 24(1):89–110, 1928.
- [117] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part III. Term Values and Intensities in Series in Optical Spectra. *P. Camb. Philos. Soc.*, 24(3):426–437, 1928.
- [118] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part IV. Further Results relating to Terms of the Optical Spectrum. *P. Camb. Philos. Soc.*, 25(3):310–314, 1929.

- [119] V. Fock. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Phys.*, 61(1):126–148, 1930.
- [120] C. C. J. Roothaan. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.*, 23:69–89, 1951.
- [121] R. G. Parr. *The Quantum Theory of Molecular Electronic Structure: A Lecture-Note and Reprint Volume*. Frontiers in Chemistry. W. A. Benjamin, 1963.
- [122] T. Koopmans. Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms. *Physica*, 1(1):104–113, 1934.
- [123] E. Engel and R. M. Dreizler. *Density Functional Theory: An Advanced Course*. Theoretical and Mathematical Physics. Springer, 2011.
- [124] B. G. Johnson, P. M. W. Gill, and J. A. Pople. Preliminary Results on the Performance of a Family of Density Functional Methods. *J. Chem. Phys.*, 97(10):7846–7848, 1992.
- [125] P. L. Fast, M. L. Sánchez, and D. G. Truhlar. Infinite Basis Limits in Electronic Structure Theory. *J. Chem. Phys.*, 111(7):2921–2926, 1999.
- [126] F. Jensen. *Introduction to Computational Chemistry*. Wiley, 2016.
- [127] T. Bredow and K. Jug. Theory and Range of Modern Semiempirical Molecular Orbital Methods. *Theor. Chem. Acc.*, 113(1):1–14, 2005.
- [128] W. Thiel. Semiempirical Quantum–Chemical Methods. *WIREs Comput. Mol. Sci.*, 4(2):145–157, 2014.
- [129] J. C. Slater. Atomic Shielding Constants. *Phys. Rev.*, 36:57–64, 1930.
- [130] J. A. Pople, D. P. Santry, and G. A. Segal. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *J. Chem. Phys.*, 43(10):S129–S135, 1965.
- [131] J. A. Pople, D. L. Beveridge, and P. A. Dobosh. Approximate Self-Consistent Molecular Orbital Theory. V. Intermediate Neglect of Differential Overlap. *J. Chem. Phys.*, 47(6):2026–2033, 1967.
- [132] J. A. Pople and G. A. Segal. Approximate Self-Consistent Molecular Orbital Theory. II. Calculations with Complete Neglect of Differential Overlap. *J. Chem. Phys.*, 43(10):S136–S151, 1965.
- [133] R. C. Bingham, M. J. S. Dewar, and D. H. Lo. Ground States of Molecules. XXV. MINDO/3. Improved Version of the MINDO Semiempirical SCF-MO Method. *J. Am. Chem. Soc.*, 97(6):1285–1293, 1975.
- [134] M. J. S. Dewar. Quantum Organic Chemistry. *Science*, 187(4181):1037–1044, 1975.

## Bibliography

---

- [135] M. J. S. Dewar and W. Thiel. Ground States of Molecules. 38. The MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.*, 99(15):4899–4907, 1977.
- [136] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.*, 107(13):3902–3909, 1985.
- [137] J. J. P. Stewart. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.*, 10(2):209–220, 1989.
- [138] M. J. S. Dewar and W. Thiel. A Semiempirical Model for the Two-Center Repulsion Integrals in the NDDO Approximation. *Theor. Chim. Acta*, 46(2):89–104, 1977.
- [139] J. J. P. Stewart. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.*, 13(12):1173–1213, 2007.
- [140] J. J. P. Stewart. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.*, 19(1):1–32, 2013.
- [141] L. H. Thomas. The Calculation of Atomic Fields. *P. Camb. Philos. Soc.*, 23(5):542–548.
- [142] E. Fermi. Un Metodo Statistico per la Determinazione di Alcune Priorieta dell'Atome. *Rend. Accad. Naz. Lincei*, 6(32):602–607, 1927.
- [143] J. C. Slater. A Simplification of the Hartree-Fock Method. *Phys. Rev.*, 81:385–390, 1951.
- [144] R. Gáspár. Über eine Approximation des Hartree-Fockschen Potentials durch eine Universelle Potentialfunktion. *Acta Phys. Acad. Sci. Hung.*, 3(3):263–286, 1954.
- [145] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136:B864–B871, 1964.
- [146] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.
- [147] M. Levy. Universal Variational Functionals of Electron Densities, First-Order Density Matrices, and Natural Spin-Orbitals and Solution of the V-Representability Problem. *P. Natl. Acad. Sci. USA*, 76(12):6062–6065, 1979.
- [148] M. Levy. Electron Densities in Search of Hamiltonians. *Phys. Rev. A*, 26:1200–1208, 1982.
- [149] E. H. Lieb. Density Functionals for Coulomb Systems. *Int. J. Quantum Chem.*, 24(3):243–277, 1983.
- [150] E. H. Lieb. *Density Functionals for Coulomb Systems*, In R. M. Dreizler and J. da Providência (editors), *Density Functional Methods In Physics*, pages 31–80. Springer, 1985.

- 
- [151] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140:A1133–A1138, 1965.
- [152] J. P. Perdew and K. Schmidt. Jacob’s Ladder of Density Functional Approximations for the Exchange-Correlation Energy. *AIP Conf. Proc.*, 577(1):1–20, 2001.
- [153] P. A. M. Dirac. Note on Exchange Phenomena in the Thomas Atom. *P. Camb. Philos. Soc.*, 26(3):376–385, 1930.
- [154] J. P. Perdew and A. Zunger. Self-Interaction Correction to Density-Functional Approximations for Many-Electron Systems. *Phys. Rev. B*, 23:5048–5079, 1981.
- [155] J. P. Perdew and Y. Wang. Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy. *Phys. Rev. B*, 45:13244–13249, 1992.
- [156] D. M. Ceperley and B. J. Alder. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.*, 45:566–569, 1980.
- [157] S. H. Vosko, L. Wilk, and M. Nusair. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.*, 58(8):1200–1211, 1980.
- [158] J. Kohanoff and N. I. Gidopoulos. *Density Functional Theory: Basics, New Trends and Applications*, In S. Wilson, P. F. Bernath, and R. McWeeny (editors), *Handbook of Molecular Physics and Quantum Chemistry, Vol. 2*, pages 532–568. Wiley, 2003.
- [159] K. Burke. The ABC of DFT. <http://www.chem.uci.edu/~kieron/dftold2/materials/bookABCDFT/gamma/g1.pdf>, 2007.
- [160] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77:3865–3868, 1996.
- [161] A. D. Becke. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A*, 38:3098–3100, 1988.
- [162] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B*, 37:785–789, 1988.
- [163] R. Colle and O. Salvetti. Approximate Calculation of the Correlation Energy for the Closed Shells. *Theor. Chim. Acta*, 37(4):329–334, 1975.
- [164] J. P. Perdew and L. A. Constantin. Laplacian-Level Density Functionals for the Kinetic Energy Density and Exchange-Correlation Energy. *Phys. Rev. B*, 75:155109, 2007.
- [165] R. T. Sharp and G. K. Horton. A Variational Approach to the Unipotential Many-Electron Problem. *Phys. Rev.*, 90:317, 1953.

## Bibliography

---

- [166] J. B. Krieger, Y. Li, and G. J. Iafrate. Construction and Application of an Accurate Local Spin-Polarized Kohn-Sham Potential with Integer Discontinuity: Exchange-Only Theory. *Phys. Rev. A*, 45:101–126, 1992.
- [167] C. Adamo, M. Ernzerhof, and G. E. Scuseria. The Meta-GGA Functional: Thermochemistry with a Kinetic Energy Density Dependent Exchange-Correlation Functional. *J. Chem. Phys.*, 112(6):2643–2649, 2000.
- [168] R. Peverati and D. G. Truhlar. Quest for a Universal Density Functional: The Accuracy of Density Functionals Across a Broad Spectrum of Databases in Chemistry and Physics. *Philos. T. Roy. Soc. A*, 372(2011), 2014.
- [169] Y. Zhao and D. G. Truhlar. A New Local Density Functional for Main-Group Thermochemistry, Transition Metal Bonding, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Phys.*, 125(19):194101, 2006.
- [170] R. Peverati and D. G. Truhlar. M11-L: A Local Density Functional that Provides Improved Accuracy for Electronic Structure Calculations in Chemistry and Physics. *J. Phys. Chem. Lett.*, 3(1):117–124, 2012.
- [171] J. Sun, A. Ruzsinszky, and J. P. Perdew. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.*, 115:036402, 2015.
- [172] J. Harris. Adiabatic-Connection Approach to Kohn-Sham Theory. *Phys. Rev. A*, 29:1648–1659, 1984.
- [173] A. D. Becke. A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.*, 98(2):1372–1377, 1993.
- [174] A. D. Becke. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.*, 98(7):5648–5652, 1993.
- [175] F. Corà, M. Alfredsson, G. Mallia, D. S. Middlemiss, W. C. Mackrodt, R. Dovesi, and R. Orlando. *The Performance of Hybrid Density Functionals in Solid State Chemistry*, In *Principles and Applications of Density Functional Theory in Inorganic Chemistry II*, pages 171–232. Springer, 2004.
- [176] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.*, 98(45):11623–11627, 1994.
- [177] G. E. Scuseria and V. N. Staroverov. Progress in the Development of Exchange-Correlation Functionals. In C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria (editors), *Theory and Application of Computational Chemistry: The First 40 Years*, pages 669–724. Elsevier, 2005.
- [178] M. Ernzerhof and G. E. Scuseria. Assessment of the Perdew-Burke-Ernzerhof Exchange-Correlation Functional. *J. Chem. Phys.*, 110(11):5029–5036, 1999.

- [179] C. Adamo and V. Barone. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.*, 110(13):6158–6170, 1999.
- [180] J. P. Perdew, M. Ernzerhof, and K. Burke. Rationale for Mixing Exact Exchange with Density Functional Approximations. *J. Chem. Phys.*, 105(22):9982–9985, 1996.
- [181] R. A. Evarestov. *Quantum Chemistry of Solids: LCAO Treatment of Crystals and Nanostructures*. Springer Series in Solid-State Sciences. Springer, 2013.
- [182] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler. Resolution-of-Identity Approach to Hartree-Fock, Hybrid Density Functionals, RPA, MP2 and GW with Numeric Atom-Centered Orbital Basis Functions. *New J. Phys.*, 14(5):053020, 2012.
- [183] Y. Zhao and D. G. Truhlar. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.*, 120(1):215–241, 2008.
- [184] Y. Zhao and D. G. Truhlar. Exploring the Limit of Accuracy of the Global Hybrid Meta Density Functional for Main-Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.*, 4(11):1849–1868, 2008.
- [185] Roberto Peverati and Donald G. Truhlar. Improving the Accuracy of Hybrid Meta-GGA Density Functionals by Range Separation. *J. Phys. Chem. Lett.*, 2(21):2810–2817, 2011.
- [186] J.-D. Chai and M. Head-Gordon. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.*, 128(8):084106, 2008.
- [187] K. Hui and J.-D. Chai. SCAN-Based Hybrid and Double-Hybrid Density Functionals From Models Without Fitted Parameters. *J. Chem. Phys.*, 144(4):044114, 2016.
- [188] D. C. Langreth and J. P. Perdew. The Exchange-Correlation Energy of a Metallic Surface. *Solid State Commun.*, 17(11):1425–1429, 1975.
- [189] O. Gunnarsson and B. I. Lundqvist. Exchange and Correlation in Atoms, Molecules, and Solids by the Spin-Density-Functional Formalism. *Phys. Rev. B*, 13:4274–4298, 1976.
- [190] D. C. Langreth and J. P. Perdew. Exchange-Correlation Energy of a Metallic Surface: Wave-Vector Analysis. *Phys. Rev. B*, 15:2884–2901, 1977.
- [191] A. Görling and M. Levy. Correlation-Energy Functional and Its High-Density Limit Obtained From a Coupling-Constant Perturbation Expansion. *Phys. Rev. B*, 47:13105–13113, 1993.
- [192] A. Görling and M. Levy. Exact Kohn-Sham Scheme Based on Perturbation Theory. *Phys. Rev. A*, 50:196–204, 1994.

## Bibliography

---

- [193] E. Engel. *Orbital-Dependent Functionals for the Exchange-Correlation Energy: A Third Generation of Density Functionals*, In C. Fiolhais, F. Nogueira, and M. A. L. Marques (editors), *A Primer in Density Functional Theory*, pages 56–122. Springer, 2003.
- [194] Y. Zhao, B. J. Lynch, and D. G. Truhlar. Doubly Hybrid Meta DFT: New Multi-Coefficient Correlation and Density Functional Methods for Thermochemistry and Thermochemical Kinetics. *J. Phys. Chem. A*, 108(21):4786–4791, 2004.
- [195] S. Grimme. Semiempirical Hybrid Density Functional with Perturbative Second-Order Correlation. *J. Chem. Phys.*, 124(3):034108, 2006.
- [196] F. Neese, T. Schwabe, and S. Grimme. Analytic Derivatives for Perturbatively Corrected "Double Hybrid" Density Functionals: Theory, Implementation, and Applications. *J. Chem. Phys.*, 126(12):124115, 2007.
- [197] Y. Zhang, X. Xu, and W. A. Goddard. Doubly Hybrid Density Functional for Accurate Descriptions of Nonbond Interactions, Thermochemistry, and Thermochemical Kinetics. *Proc. Natl. Acad. Sci. USA*, 106(13):4963–4968, 2009.
- [198] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople. Assessment of Gaussian-3 and Density Functional Theories for a Larger Experimental Test Set. *J. Chem. Phys.*, 112(17):7374–7383, 2000.
- [199] S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen. Consistent Structures and Interactions by Density Functional Theory with Small Atomic Orbital Basis Sets. *J. Chem. Phys.*, 143(5):054107, 2015.
- [200] F. Weigend and R. Ahlrichs. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.*, 7:3297–3305, 2005.
- [201] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.*, 132(15):154104, 2010.
- [202] H. Kruse and S. Grimme. A Geometrical Correction for the Inter- and Intra-Molecular Basis Set Superposition Error in Hartree-Fock and Density Functional Theory Calculations for Large Systems. *J. Chem. Phys.*, 136(15):154101, 2012.
- [203] H. B. Jansen and P. Ros. Non-Empirical Molecular Orbital Calculations on the Protonation of Carbon Monoxide. *Chem. Phys. Lett.*, 3(3):140–143, 1969.
- [204] B. Liu and A. D. McLean. Accurate Calculation of the Attractive Interaction of Two Ground State Helium Atoms. *J. Chem. Phys.*, 59(8):4557–4558, 1973.
- [205] S. Grimme. Density Functional Theory with London Dispersion Corrections. *WIREs Comput. Mol. Sci.*, 1(2):211–228, 2011.



- [206] M. Ropo, M. Schneider, C. Baldauf, and V. Blum. First-Principles Data Set of 45,892 Isolated and Cation-Coordinated Conformers of 20 Proteinogenic Amino Acids. *Sci. Data*, 3:160009, 2016.
- [207] F. Schubert, M. Rossi, C. Baldauf, K. Pagel, S. Warnke, G. von Helden, F. Filsinger, P. Kupser, G. Meijer, M. Salwiczek, B. Koksich, M. Scheffler, and V. Blum. Exploring the Conformational Preferences of 20-Residue Peptides in Isolation: Ac-Ala<sub>19</sub>-Lys + H<sup>+</sup> vs. Ac-Lys-Ala<sub>19</sub> + H<sup>+</sup> and the Current Reach of DFT. *Phys. Chem. Chem. Phys.*, 17:7373–7385, 2015.
- [208] A. Otero de la Roza and G. A. DiLabio. *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*. Elsevier Science, 2017.
- [209] G. A. DiLabio and A. Otero de la Roza. *Noncovalent Interactions in Density Functional Theory*, In A. L. Parrill and K. B. Lipkowitz (editors), *Reviews in Computational Chemistry*, pages 1–97. Wiley, 2016.
- [210] W. J. Deal. The Long-Range Interaction Between Two Hydrogen Atoms. *Int. J. Quantum Chem.*, 6(3):593–596, 1972.
- [211] R. Ahlrichs, R. Penco, and G. Scoles. Intermolecular Forces in Simple Systems. *Chem. Phys.*, 19(2):119–130, 1977.
- [212] E. R. Johnson and A. D. Becke. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *J. Chem. Phys.*, 124(17):174104, 2006.
- [213] A. Tkatchenko and M. Scheffler. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.*, 102:073005, 2009.
- [214] A. D. Becke and E. R. Johnson. A Density-Functional Model of the Dispersion Interaction. *J. Chem. Phys.*, 123(15):154101, 2005.
- [215] E. R. Johnson and A. D. Becke. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *J. Chem. Phys.*, 124(17):174104, 2006.
- [216] D. G. A. Smith, L. A. Burns, K. Patkowski, and C. D. Sherrill. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.*, 7(12):2197–2203, 2016.
- [217] J.-D. Chai and M. Head-Gordon. Long-Range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.*, 10:6615–6620, 2008.
- [218] H. B. G. Casimir and D. Polder. The Influence of Retardation on the London-van der Waals Forces. *Phys. Rev.*, 73:360–372, 1948.

## Bibliography

---

- [219] B. M. Axilrod and E. Teller. Interaction of the van der Waals Type Between Three Atoms. *J. Chem. Phys.*, 11(6):299–300, 1943.
- [220] Y. Muto. Force Between Nonpolar Molecules. *Proc. Phys. Math. Soc. Jpn.*, 17:629–631, 1943.
- [221] Q. Wu and W. Yang. Empirical Correction to Density Functional Theory for van der Waals Interactions. *J. Chem. Phys.*, 116(2):515–524, 2002.
- [222] P. Jurečka, J. Šponer, J. Černý, and P. Hobza. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.*, 8:1985–1993, 2006.
- [223] F. L. Hirshfeld. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta*, 44(2):129–138, 1977.
- [224] E. R. Johnson and A. D. Becke. A Post-Hartree-Fock Model of Intermolecular Interactions. *J. Chem. Phys.*, 123(2):024101, 2005.
- [225] A. Olasz, K. Vanommeslaeghe, A. Krishtal, T. Veszprémi, C. Van Alsenoy, and P. Geerlings. The Use of Atomic Intrinsic Polarizabilities in the Evaluation of the Dispersion Energy. *J. Chem. Phys.*, 127(22):224105, 2007.
- [226] K. T. Tang. Dynamic Polarizabilities and van der Waals Coefficients. *Phys. Rev.*, 177:108–114, 1969.
- [227] X. Chu and A. Dalgarno. Linear Response Time-Dependent Density Functional Theory for van der Waals Coefficients. *J. Chem. Phys.*, 121(9):4083–4088, 2004.
- [228] A. Ambrosetti, A. M. Reilly, R. A. DiStasio Jr., and A. Tkatchenko. Long-Range Correlation Energy Calculated from Coupled Atomic Response Functions. *J. Chem. Phys.*, 140(18):18A508, 2014.
- [229] A. Tkatchenko, R. A. DiStasio Jr., R. Car, and M. Scheffler. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.*, 108:236402, 2012.
- [230] A. Tkatchenko, A. Ambrosetti, and R. A. DiStasio Jr. Interatomic Methods for the Dispersion Energy Derived from the Adiabatic Connection Fluctuation-Dissipation Theorem. *J. Chem. Phys.*, 138(7):074106, 2013.
- [231] J. Řezáč and P. Hobza. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.*, 8(1):141–151, 2012.
- [232] J. Řezáč, K. E. Riley, and P. Hobza. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.*, 7(8):2427–2438, 2011.

- [233] B. Vorlová, D. Nachtigallová, J. Jirásková-Vaníčková, H. Ajani, P. Jansa, J. Řezáč, J. Fanfrlík, M. Otyepka, P. Hobza, J. Konvalinka, and M. Lepšík. Malonate-Based Inhibitors of Mammalian Serine Racemase: Kinetic Characterization and Structure-Based Computational Study. *Eur. J. Med. Chem.*, 89:189–197, 2015.
- [234] V. Magnasco. *Elementary Methods of Molecular Quantum Mechanics*. Elsevier Science, 2006.
- [235] E. G. Lewars. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. Springer, 2016.
- [236] A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education. Prentice Hall, 2001.
- [237] S. Grabowski. *Hydrogen Bonding - New Insights*. Challenges and Advances in Computational Chemistry and Physics. Springer, 2006.
- [238] E. Schrödinger. Quantisierung als Eigenwertproblem. *Ann. Phys.*, 385(13):437–490, 1926.
- [239] J. W. S. Rayleigh. *The Theory of Sound (Republication of the 1894 Second Edition) – Vol. 1*. Macmillan, 1944.
- [240] C. Møller and M. S. Plesset. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.*, 46:618–622, 1934.
- [241] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Books on Chemistry. Dover Publications, 2012.
- [242] P. Hobza and K. Müller-Dethlefs. *Non-Covalent Interactions: Theory and Experiment*. RSC Theoretical and Computational Chemistry Series. The Royal Society of Chemistry, 2009.
- [243] J. Olsen, P. Jørgensen, T. Helgaker, and O. Christiansen. Divergence in Møller–Plesset Theory: A Simple Explanation Based on a Two-State Model. *J. Chem. Phys.*, 112(22):9736–9748, 2000.
- [244] E. R. Davidson. *Perspectives on Ab Initio Calculations*, In K. B. Lipkowitz and D. B. Boyd (editors), *Reviews in Computational Chemistry, Volume 1*, pages 373–382. Wiley, 2007.
- [245] F. Coester and H. Kümmel. Short-Range Correlations in Nuclear Wave Functions. *Nucl. Phys.*, 17(Supplement C):477–485, 1960.
- [246] J. Čížek. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *J. Chem. Phys.*, 45(11):4256–4266, 1966.

## Bibliography

---

- [247] J. Čížek. *On the Use of the Cluster Expansion and the Technique of Diagrams in Calculations of Correlation Effects in Atoms and Molecules*, In R. LeFebvre and C. Moser (editors), *Advances in Chemical Physics: Correlation Effects in Atoms and Molecules, Volume 14*, pages 35–89. Wiley, 1969.
- [248] G. D. Purvis III and R. J. Bartlett. A Full Coupled-Cluster Singles and Doubles Model: The Inclusion of Disconnected Triples. *J. Chem. Phys.*, 76(4):1910–1918, 1982.
- [249] J. D. Watts and R. J. Bartlett. Triple Excitations in Coupled-Cluster Theory: Energies and Analytical Derivatives. *Int. J. Quantum Chem.*, 48(S27):51–66, 1993.
- [250] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon. A Fifth-Order Perturbation Comparison of Electron Correlation Theories. *Chem. Phys. Lett.*, 157(6):479–483, 1989.
- [251] K. E. Riley, M. Pitoňák, P. Jurečka, and P. Hobza. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.*, 110(9):5023–5063, 2010.
- [252] J. Řezáč and P. Hobza. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the “Gold Standard” CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.*, 9(5):2151–2155, 2013.
- [253] C. Riplinger and F. Neese. An Efficient and Near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.*, 138(3):034106, 2013.
- [254] C. Riplinger, B. Sandhoefer, A. Hansen, and F. Neese. Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *J. Chem. Phys.*, 139(13):134101, 2013.
- [255] H.-J. Werner, F. R. Manby, and P. J. Knowles. Fast Linear Scaling Second-Order Møller-Plesset Perturbation Theory (MP2) Using Local and Density Fitting Approximations. *J. Chem. Phys.*, 118(18):8149–8160, 2003.
- [256] J. W. Boughton and P. Pulay. Comparison of the Boys and Pipek-Mezey Localizations in the Local Correlation Approach and Automatic Virtual Basis Selection. *J. Comput. Chem.*, 14(6):736–740, 1993.
- [257] C. Edmiston and M. Krauss. Configuration-Interaction Calculation of H<sub>3</sub> and H<sub>2</sub>. *J. Chem. Phys.*, 42(3):1119–1120, 1965.
- [258] C. Edmiston and M. Krauss. Pseudonatural Orbitals as a Basis for the Superposition of Configurations. I. He<sub>2</sub><sup>+</sup>. *J. Chem. Phys.*, 45(5):1833–1839, 1966.
- [259] W. Meyer. *Configuration Expansion by Means of Pseudonatural Orbitals*, In H. F. Schaefer (editor), *Methods of Electronic Structure Theory*, pages 413–446. Springer, 1977.

- [260] J. W. Ponder. TINKER - Software Tools for Molecular Design. *Washington University School of Medicine, Saint Louis, MO*, 2013.
- [261] J. J. P. Stewart. MOPAC2016. Stewart Computational Chemistry, <http://OpenMOPAC.net>.
- [262] J. J. P. Stewart. *Semiempirical Molecular Orbital Methods*, In *Reviews in Computational Chemistry, Volume 1*, pages 45–81. Wiley, 2007.
- [263] J. J. P. Stewart. MOPAC2016 Manual. [http://openmopac.net/manual/SCF\\_calc\\_hof.html](http://openmopac.net/manual/SCF_calc_hof.html).
- [264] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler. Ab Initio Molecular Simulations with Numeric Atom-Centered Orbitals. *Comput. Phys. Commun.*, 180(11):2175–2196, 2009.
- [265] T. Auckenthaler, V. Blum, H.-J. Bungartz, T. Huckle, R. Johanni, L. Krämer, B. Lang, H. Lederer, and P. R. Willems. Parallel Solution of Partial Symmetric Eigenvalue Problems from Electronic Structure Calculations. *Parallel Comput.*, 37(12):783–794, 2011.
- [266] E. van Lenthe, E. J. Baerends, and J. G. Snijders. Relativistic Total Energy Using Regular Approximations. *J. Chem. Phys.*, 101(11):9783–9792, 1994.
- [267] I. Y. Zhang, X. Ren, P. Rinke, V. Blum, and M. Scheffler. Numeric Atom-Centered-Orbital Basis Sets with Valence-Correlation Consistency From H to Ar. *New J. Phys.*, 15(12):123033, 2013.
- [268] T. H. Dunning Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.*, 90(2):1007–1023, 1989.
- [269] N. B. Balabanov and K. A. Peterson. Systematically Convergent Basis Sets for Transition Metals. I. All-Electron Correlation Consistent Basis Sets for the 3d Elements Sc-Zn. *J. Chem. Phys.*, 123(6):064107, 2005.
- [270] I. Y. Zhang, Y. Luo, and X. Xu. Basis Set Dependence of the Doubly Hybrid XYG3 Functional. *J. Chem. Phys.*, 133(10):104105, 2010.
- [271] F. Neese. The ORCA Program System. *WIREs Comput. Mol. Sci.*, 2(1):73–78, 2012.
- [272] E. van Lenthe, E. J. Baerends, and J. G. Snijders. Relativistic Regular Two-Component Hamiltonians. *J. Chem. Phys.*, 99(6):4597–4610, 1993.
- [273] C. van Wüllen. Molecular Density Functional Calculations in the Regular Relativistic Approximation: Method, Application to Coinage Metal Diatomics, Hydrides, Fluorides and Chlorides, and Comparison with First-Order Relativistic Calculations. *J. Chem. Phys.*, 109(2):392–399, 1998.
- [274] D. A. Pantazis, X.-Y. Chen, C. R. Landis, and F. Neese. All-Electron Scalar Relativistic Basis Sets for Third-Row Transition Metal Atoms. *J. Chem. Theory Comput.*, 4(6):908–919, 2008.

## Bibliography

---

- [275] D. G. Truhlar. Basis-Set Extrapolation. *Chem. Phys. Lett.*, 294(1):45–48, 1998.
- [276] A. Karton and J. M. L. Martin. Comment on: “Estimating the Hartree-Fock limit from finite basis set calculations” [Jensen F. (2005) *Theor. Chem. Acc.* 113:267]. *Theor. Chem. Acc.*, 115(4):330–333, 2006.
- [277] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson. Basis-Set Convergence in Correlated Calculations on Ne, N<sub>2</sub>, and H<sub>2</sub>O. *Chem. Phys. Lett.*, 286(3):243–252, 1998.
- [278] S. F. Boys and F. Bernardi. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures With Reduced Errors. *Mol. Phys.*, 19(4):553–566, 1970.
- [279] C. A. Floudas and C. E. Gounaris. A Review of Recent Advances in Global Optimization. *Journal Global Optim.*, 45(1):3, 2008.
- [280] J. Lee, P. L. Freddolino, and Y. Zhang. *Ab Initio Protein Structure Prediction*, In D. J. Rigden (editor), *From Protein Structure to Function with Bioinformatics*, pages 3–35. Springer, 2017.
- [281] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.
- [282] B. A. Berg and T. Neuhaus. Multicanonical Ensemble: A New Approach to Simulate First-Order Phase Transitions. *Phys. Rev. Lett.*, 68:9–12, 1992.
- [283] J. Lee. New Monte Carlo Algorithm: Entropic Sampling. *Phys. Rev. Lett.*, 71:211–214, 1993.
- [284] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.
- [285] Y. Zhang, D. Kihara, and J. Skolnick. Local Energy Landscape Flattening: Parallel Hyperbolic Monte Carlo Sampling of Protein Folding. *Proteins: Struct., Funct., Bioinf.*, 48(2):192–201, 2002.
- [286] Y. Sugita and Y. Okamoto. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.*, 314(1):141–151, 1999.
- [287] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.*, 120(24):11919–11929, 2004.
- [288] A. Supady, V. Blum, and C. Baldauf. First-Principles Molecular Structure Search with a Genetic Algorithm. *J. Chem. Inf. Model.*, 55(11):2338–2348, 2015.

- [289] J. Lee, H. A. Scheraga, and S. Rackovsky. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers*, 46(2):103–115, 1998.
- [290] Z. Li and H. A. Scheraga. Monte Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding. *Proc. Natl. Acad. Sci. USA*, 84(19):6611–6615, 1987.
- [291] D. J. Wales and J. P. K. Doye. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A*, 101(28):5111–5116, 1997.
- [292] R. V. Pappu, R. K. Hart, and J. W. Ponder. Hierarchical Conformational Scanning Using Potential Smoothing. Application to Cycloheptadecane. <https://dasher.wustl.edu/ponder/papers/ccb-report-1998-01.pdf>, 1998.
- [293] J. W. Ponder and F. M. Richards. An Efficient Newton-Like Method for Molecular Mechanics Energy Minimization of Large Molecules. *J. Comput. Chem.*, 8(7):1016–1024, 1987.
- [294] T. D. J. Perkins. *Exploiting Similarity Between Highly Flexible and Dissimilar Molecular Structures*, In P. M. Dean (editor), *Molecular Similarity in Drug Design*, pages 89–109. Springer, 1995.
- [295] H. Haken and H. C. Wolf. *Molecular Physics and Elements of Quantum Chemistry: Introduction to Experiments and Theory*. Advanced Texts in Physics. Springer, 2004.
- [296] M. Wolfsberg, W.A. Van Hook, P. Paneth, and L. P. N. Rebelo. *Isotope Effects: in the Chemical, Geological, and Bio Sciences*. Springer, 2009.
- [297] E. B. Wilson Jr., J. C. Decius, and P. C. Cross. *Molecular Vibrations: The Theory of Infrared and Vibrational Spectra*. McGraw-Hill, 1955.
- [298] E. Schrödinger. Der stetige Übergang von der Mikro- zur Makromechanik. *Naturwissenschaften*, 14(28):664–666, 1926.
- [299] L. Pauling and E. B. Wilson. *Introduction to Quantum Mechanics with Applications to Chemistry*. Dover Books on Physics. Dover Publications, 1935.
- [300] P. A. M. Dirac. The Quantum Theory of the Emission and Absorption of Radiation. *P. Roy. Soc. Lond. A Mat.*, 114(767):243–265, 1927.
- [301] E. Fermi. *Nuclear Physics: A Course Given by Enrico Fermi at the University of Chicago*. 1950.
- [302] H. P. Figeys and P. Geerlings. *Some Aspects of the Quantumchemical Interpretation of Integrated Intensities of Infrared Absorption Bands*, In Z. B. Maksić (editor), *Theoretical Models of Chemical Bonding, Part 3: Molecular Spectroscopy, Electronic Structure and Intramolecular Interactions*, pages 25–62. Springer, 1991.

## Bibliography

---

- [303] D. R. Yarkony. *Modern Electronic Structure Theory, Part 1*. Advanced Series in Physical Chemistry. World Scientific, 1995.
- [304] J. Neugebauer, M. Reiher, C. Kind, and B. A. Hess. Quantum Chemical Calculation of Vibrational Spectra of Large Molecules – Raman and IR Spectra for Buckminsterfullerene. *J. Comput. Chem.*, 23(9):895–910, 2002.
- [305] D. Porezag and M. R. Pederson. Infrared Intensities and Raman-Scattering Activities within Density-Functional Theory. *Phys. Rev. B*, 54:7830–7836, 1996.
- [306] R. Gehrke. *First-Principles Basin-Hopping for the Structure Determination of Atomic Clusters*. PhD thesis, Freie Universität Berlin (FU Berlin), 2009.
- [307] R. F. W. Bader, A. Larouche, C. Gatti, M. T. Carroll, P. J. MacDougall, and K. B. Wiberg. Properties of Atoms in Molecules: Dipole Moments and Transferability of Properties. *J. Chem. Phys.*, 87(2):1142–1152, 1987.
- [308] D. A. McQuarrie. *Statistical Mechanics*. University Science Books, 2000.
- [309] T. L. Hill. *An Introduction to Statistical Thermodynamics*. Addison-Wesley Series in Chemistry. Dover Publications, 1960.
- [310] I. N. Levine. *Physical Chemistry*. McGraw-Hill, 1988.
- [311] K. Lucas. *Applied Statistical Thermodynamics*. Springer, 2013.
- [312] R. F. Sekerka. *Thermal Physics: Thermodynamics and Statistical Mechanics for Scientists and Engineers*. Elsevier Science, 2015.
- [313] A. Svendsen, U. J. Lorenz, O. V. Boyarkin, and T. R. Rizzo. A New Tandem Mass Spectrometer for Photofragment Spectroscopy of Cold, Gas-Phase Molecular Ions. *Rev. Sci. Instrum.*, 81(7):073107, 2010.
- [314] L. Voronina. *Gas-Phase Probes of Kinetically Trapped Peptides*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2016.
- [315] N. S. Nagornova, T. R. Rizzo, and O. V. Boyarkin. Exploring the Mechanism of IR-UV Double-Resonance for Quantitative Spectroscopy of Protonated Polypeptides and Proteins. *Angew. Chem. Int. Edit.*, 52(23):6002–6005, 2013.
- [316] M. Schneider, C. Masellis, T. Rizzo, and C. Baldauf. Kinetically Trapped Liquid-State Conformers of a Sodiated Model Peptide Observed in the Gas Phase. *J. Phys. Chem. A*, 121(36):6838–6844, 2017.
- [317] D. J. Barlow and J. M. Thornton. Helix Geometry in Proteins. *J. Mol. Biol.*, 201(3):601–619, 1988.
- [318] R. R. Hudgins, M. A. Ratner, and M. F. Jarrold. Design of Helices That Are Stable in Vacuo. *J. Am. Chem. Soc.*, 120(49):12974–12975, 1998.



- [319] R. R. Hudgins and M. F. Jarrold. Helix Formation in Unsolvated Alanine-Based Peptides: Helical Monomers and Helical Dimers. *J. Am. Chem. Soc.*, 121(14):3494–3501, 1999.
- [320] M. Kohtani and M. F. Jarrold. Water Molecule Adsorption on Short Alanine Peptides: How Short Is the Shortest Gas-Phase Alanine-Based Helix? *J. Am. Chem. Soc.*, 126(27):8454–8458, 2004.
- [321] W. Chin, F. Piuzzi, J.-P. Dognon, I. Dimicoli, B. Tardivel, and M. Mons. Gas Phase Formation of a  $3_{10}$ -Helix in a Three-Residue Peptide Chain: Role of Side Chain-Backbone Interactions as Evidenced by IR-UV Double Resonance Experiments. *J. Am. Chem. Soc.*, 127(34):11900–11901, 2005.
- [322] J. A. Stearns, O. V. Boyarkin, and T. R. Rizzo. Spectroscopic Signatures of Gas-Phase Helices: Ac-Phe-(Ala)<sub>5</sub>-Lys-H<sup>+</sup> and Ac-Phe-(Ala)<sub>10</sub>-Lys-H<sup>+</sup>. *J. Am. Chem. Soc.*, 129(45):13820–13821, 2007.
- [323] J. A. Stearns, O. V. Boyarkin, and T. R. Rizzo. Effects of N-Terminus Substitution on the Structure and Spectroscopy of Gas-Phase Helices. *CHIMIA*, 62(4):240–243, 2008.
- [324] J. A. Stearns, C. Seaiby, O. V. Boyarkin, and T. R. Rizzo. Spectroscopy and Conformational Preferences of Gas-Phase Helices. *Phys. Chem. Chem. Phys.*, 11:125–132, 2009.
- [325] A. V. Zabuga and T. R. Rizzo. Capping Motif for Peptide Helix Formation. *J. Phys. Chem. Lett.*, 6(9):1504–1508, 2015.
- [326] M. Rossi, V. Blum, P. Kupser, G. von Helden, F. Bierau, K. Pagel, G. Meijer, and M. Scheffler. Secondary Structure of Ac-Ala<sub>n</sub>-LysH<sup>+</sup> Polyalanine Peptides (n = 5,10,15) in Vacuo: Helical or Not? *J. Phys. Chem. Lett.*, 1(24):3465–3470, 2010.
- [327] F. Schubert, K. Pagel, M. Rossi, S. Warnke, M. Salwiczek, B. Koksich, G. von Helden, V. Blum, C. Baldauf, and M. Scheffler. Native Like Helices in a Specially Designed  $\beta$  Peptide in the Gas Phase. *Phys. Chem. Chem. Phys.*, 17:5376–5385, 2015.
- [328] F. Schubert, M. Rossi, C. Baldauf, K. Pagel, S. Warnke, G. von Helden, F. Filsinger, P. Kupser, G. Meijer, M. Salwiczek, B. Koksich, M. Scheffler, and V. Blum. Exploring the Conformational Preferences of 20-Residue Peptides in Isolation: Ac-Ala<sub>19</sub>-Lys + H<sup>+</sup> vs. Ac-Lys-Ala<sub>19</sub> + H<sup>+</sup> and the Current Reach of DFT. *Phys. Chem. Chem. Phys.*, 17:7373–7385, 2015.
- [329] C. Baldauf and M. Rossi. Going Clean: Structure and Dynamics of Peptides in the Gas Phase and Paths to Solvation. *J. Phys.: Condens. Mat.*, 27(49):493002, 2015.
- [330] M. Rossi, M. Scheffler, and V. Blum. Impact of Vibrational Entropy on the Stability of Unsolvated Peptide Helices with Increasing Length. *J. Phys. Chem. B*, 117(18):5574–5584, 2013.

## Bibliography

---

- [331] C. Baldauf and H.-J. Hofmann. Ab Initio MO Theory - An Important Tool in Foldamer Research: Prediction of Helices in Oligomers of  $\omega$ -Amino Acids. *Helv. Chim. Acta*, 95(12):2348–2383, 2012.
- [332] M. Rossi, S. Chutia, M. Scheffler, and V. Blum. Validation Challenge of Density-Functional Theory for Peptides - Example of Ac-Phe-Ala<sub>5</sub>-LysH<sup>+</sup>. *J. Phys. Chem. A*, 118(35):7349–7359, 2014.
- [333] W. Hoffmann, M. Marianski, S. Warnke, J. Seo, C. Baldauf, G. von Helden, and K. Pagel. Assessing the Stability of Alanine-Based Helices by Conformer-Selective IR Spectroscopy. *Phys. Chem. Chem. Phys.*, 18:19950–19954, 2016.
- [334] C. Baldauf, K. Pagel, S. Warnke, G. von Helden, B. Koks, V. Blum, and M. Scheffler. How Cations Change Peptide Structure. *Chem. Eur. J.*, 19(34):11224–11234, 2013.
- [335] M. Ropo, V. Blum, and C. Baldauf. Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead. *Sci. Rep.*, 6:35772, 2016.
- [336] S. De, F. Musil, T. Ingram, C. Baldauf, and M. Ceriotti. Mapping and classifying molecules from a high-throughput structural database. *J. Cheminform.*, 9(1):6, 2017.
- [337] M. Kohtani, B. S. Kinnear, and M. F. Jarrold. Metal-Ion Enhanced Helicity in the Gas Phase. *J. Am. Chem. Soc.*, 122(49):12377–12378, 2000.
- [338] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.
- [339] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An Open Chemical Toolbox. *J. Cheminform.*, 3(1):33, 2011.
- [340] R. Sokal and C. Michener. A Statistical Method for Evaluating Systematic Relationships. *Univ. Kans. Sci. Bull.*, 38:1409–1438, 1958.
- [341] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open Source Scientific Tools for Python, 2001–.
- [342] A. P. Scott and L. Radom. Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Møller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.*, 100(41):16502–16513, 1996.
- [343] J. P. Merrick, D. Moran, and L. Radom. An Evaluation of Harmonic Vibrational Frequency Scale Factors. *J. Phys. Chem. A*, 111(45):11683–11700, 2007.
- [344] I. M. Alecu, J. Zheng, Y. Zhao, and D. G. Truhlar. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.*, 6(9):2872–2887, 2010.

- [345] I. Borukhov, D. Andelman, and H. Orland. Adsorption of Large Ions from an Electrolyte Solution: A Modified Poisson-Boltzmann Equation. *Electrochim. Acta*, 46(2):221–229, 2000.
- [346] I. Borukhov, D. Andelman, and H. Orland. Steric Effects in Electrolytes: A Modified Poisson-Boltzmann Equation. *Phys. Rev. Lett.*, 79:435–438, 1997.
- [347] S. Ringe, H. Oberhofer, C. Hille, S. Matera, and K. Reuter. Function-Space-Based Solution Scheme for the Size-Modified Poisson-Boltzmann Equation in Full-Potential DFT. *J. Chem. Theory Comput.*, 12(8):4052–4066, 2016.
- [348] L. Voronina, A. Masson, M. Kamrath, F. Schubert, D. Clemmer, C. Baldauf, and T. Rizzo. Conformations of Prolyl-Peptide Bonds in the Bradykinin 1-5 Fragment in Solution and in the Gas Phase. *J. Am. Chem. Soc.*, 138(29):9224–9233, 2016.
- [349] M. Schneider and C. Baldauf. Relative Energetics of Acetyl-Histidine Protomers with and without  $\text{Zn}^{2+}$  and a Benchmark of Energy Methods. *to be submitted*, 2018.
- [350] E. A. Amin and D. G. Truhlar. Zn Coordination Chemistry: Development of Benchmark Suites for Geometries, Dipole Moments, and Bond Dissociation Energies and Their Use To Test and Validate Density Functionals and Molecular Orbital Theory. *J. Chem. Theory Comput.*, 4(1):75–85, 2008.
- [351] V. M. Rayón, H. Valdés, N. D[í]az, and D. Suárez. Monoligand Zn(II) Complexes: Ab Initio Benchmark Calculations and Comparison with Density Functional Theory Methodologies. *J. Chem. Theory Comput.*, 4(2):243–256, 2008.
- [352] M. N. Weaver, K. M. Merz, D. Ma, H. J. Kim, and L. Gagliardi. Calculation of Heats of Formation for Zn Complexes: Comparison of Density Functional Theory, Second Order Perturbation Theory, Coupled-Cluster and Complete Active Space Methods. *J. Chem. Theory Comput.*, 9(12):5277–5285, 2013.
- [353] O. Gutten, I. Bešševová, and L. Rulíšek. Interaction of Metal Ions with Biomolecular Ligands: How Accurate Are Calculated Free Energies Associated with Metal Ion Complexation? *J. Phys. Chem. A*, 115(41):11394–11402, 2011.
- [354] O. Gutten and L. Rulíšek. Predicting the Stability Constants of Metal-Ion Complexes from First Principles. *Inorg. Chem.*, 52(18):10347–10355, 2013.
- [355] V. Navrátil, V. Klusák, and L. Rulíšek. Theoretical Aspects of Hydrolysis of Peptide Bonds by Zinc Metalloenzymes. *Chem. Eur. J.*, 19(49):16634–16645.
- [356] T. Dudev and C. Lim. Competition among Metal Ions for Protein Binding Sites: Determinants of Metal Ion Selectivity in Proteins. *Chem. Rev.*, 114(1):538–556, 2014.
- [357] T. Dudev and C. Lim. Modeling  $\text{Zn}^{2+}$ -Cysteinate Complexes in Proteins. *J. Phys. Chem. B*, 105(43):10709–10714, 2001.

## Bibliography

---

- [358] T. Dudev and C. Lim. Metal-Binding Affinity and Selectivity of Nonstandard Natural Amino Acid Residues from DFT/CDM Calculations. *J. Phys. Chem. B*, 113(34):11754–11764, 2009.
- [359] O. Gutten and L. Rulíšek. How simple is too simple? Computational perspective on importance of second-shell environment for metal-ion selectivity. *Phys. Chem. Chem. Phys.*, 17:14393–14404, 2015.
- [360] D. G. Liakos, M. Sparta, M. K. Kesharwani, J. M. L. Martin, and F. Neese. Exploring the Accuracy Limits of Local Pair Natural Orbital Coupled-Cluster Theory. *J. Chem. Theory Comput.*, 11(4):1525–1539, 2015.
- [361] S. Kossmann and F. Neese. Comparison of Two Efficient Approximate Hartree-Fock Approaches. *Chem. Phys. Lett.*, 481(4):240–243, 2009.
- [362] R. Strange, F. R. Manby, and P. J. Knowles. Automatic Code Generation in Density Functional Theory. *Comput. Phys. Commun.*, 136(3):310–318, 2001.
- [363] Stefan Grimme. DFT-D3. <https://www.chemie.uni-bonn.de/pctc/mulliken-center/software/dft-d3/dft-d3>.
- [364] J. Åqvist. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.*, 94(21):8021–8024, 1990.
- [365] R. H. Stote and M. Karplus. Zinc Binding in Proteins and Solution: A Simple but Accurate Nonbonded Representation. *Proteins: Struct., Funct., Bioinf.*, 23(1):12–31, 1995.
- [366] P. Li and K. M. Merz. Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.*, 117(3):1564–1686, 2017.
- [367] M. Hülsmann, K. N. Kirschner, A. Krämer, D. D. Heinrich, O. Krämer-Fuhrmann, and D. Reith. *Optimizing Molecular Models Through Force-Field Parameterization via the Efficient Combination of Modular Program Packages*, In R. Q. Snurr, C. S. Adjiman, and D. A. Kofke (editors), *Foundations of Molecular Modeling and Simulation: Select Papers from FOMMS 2015*, pages 53–77. Springer, 2016.
- [368] L. Huang and B. Roux. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.*, 9(8):3543–3556, 2013.
- [369] L. Huang and B. Roux. General Automated Atomic Model Parameterization. <http://gaamp.lcrc.anl.gov/>, 2013.
- [370] Y. Li, H. Li, F. C. Pickard, B. Narayanan, F. G. Sen, M. K. Y. Chan, S. K. R. S. Sankaranarayanan, B. R. Brooks, and B. Roux. Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.*, 13(9):4492–4503, 2017.

- [371] E. Fracchia, G. Del Frate, G. Mancini, W. Rocchia, and V. Barone. Force Field Parametrization of Metal Ions from Statistical Learning Techniques. *J. Chem. Theory Comput.*, 14(1):255–273, 2018.
- [372] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.*, 108:058301, 2012.
- [373] M. Ceriotti, G. A. Tribello, and M. Parrinello. Simplifying the Representation of Complex Free-Energy Landscapes using Sketch-Map. *Proc. Natl. Acad. Sci. USA*, 108(32):13023–13028, 2011.
- [374] F. Comitani, K. Rossi, M. Ceriotti, M. E. Sanz, and C. Molteni. Mapping the Conformational Free Energy of Aspartic Acid in the Gas Phase and in Aqueous Solution. *J. Chem. Phys.*, 146(14):145102, 2017.
- [375] S. De, F. Musil, T. Ingram, C. Baldauf, and M. Ceriotti. Mapping and Classifying Molecules from a High-Throughput Structural Database. *J. Cheminform.*, 9(1):6, 2017.
- [376] J. Behler and M. Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98:146401, 2007.
- [377] J. Behler. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.*, 134(7):074106, 2011.
- [378] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.*, 104:136403, 2010.
- [379] A. P. Bartók, R. Kondor, and G. Csányi. On Representing Chemical Environments. *Phys. Rev. B*, 87:184115, 2013.
- [380] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani. Comparison of Permutationally Invariant Polynomials, Neural Networks, and Gaussian Approximation Potentials in Representing Water Interactions Through Many-Body Expansions. *J. Chem. Phys.*, 148(24):241725, 2018.
- [381] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.*, 148(24):241730, 2018.
- [382] C. Baldauf, K. Pagel, S. Warnke, G. von Helden, B. Koksich, V. Blum, and M. Scheffler. How Cations Change Peptide Structure. *Chem. Eur. J.*, 19(34):11224–11234, 2013.
- [383] A. M. Legendre. *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Nineteenth Century Collections Online (NCCO): Science, Technology, and Medicine: 1780-1925. F. Didot, 1805.

## Bibliography

---

- [384] S. M. Stigler. Gauss and the Invention of Least Squares. *Ann. Stat.*, 9(3):465–474, 1981.
- [385] W. W. Piegorsch. *Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery*. Wiley, 2015.
- [386] M. H. Kutner. *Applied Linear Statistical Models*. McGraw-Hill Irwin, 2005.
- [387] J. E. Gentle. *Matrix Algebra: Theory, Computations and Applications in Statistics*. Springer Texts in Statistics. Springer, 2017.
- [388] M. A. Lukas. *Regularization Methods*, In J. S. Hunter A. El-Shaarawi (editor), *Encyclopedia of Environmetrics*. Wiley, 2006.
- [389] R. Sundberg. *Shrinkage Regression*, In J. S. Hunter A. H. El-Shaarawi (editor), *Encyclopedia of Environmetrics*. Wiley, 2006.
- [390] A. N. Tikhonov. On the Solution of Ill-Posed Problems and the Method of Regularization. *Dokl. Akad. Nauk SSSR*, 151:501–504, 1963.
- [391] D. Hughes-Hallett, A. M. Gleason, and W. G. McCallum. *Calculus: Single and Multivariable*. Wiley, 2012.
- [392] B. Clarke, E. Fokoue, and H. H. Zhang. *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics. Springer, 2009.
- [393] A. E. Hoerl, R. W. Kennard, and K. F. Baldwin. Ridge Regression: Some Simulations. *Commun. Stat.*, 4(2):105–123, 1975.
- [394] G. H. Golub, M. Heath, and G. Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223, 1979.
- [395] Y. Chen, P. Du, and Y. Wang. Variable Selection in Linear Models. *WIREs Comput. Stat.*, 6(1):1–9, 2014.
- [396] A. M. Belostotskii. *Conformational Concept For Synthetic Chemist's Use: Principles And In Lab Exploitation*. World Scientific, 2015.
- [397] F. A. Momany. Determination of Partial Atomic Charges from Ab Initio Molecular Electrostatic Potentials. Application to Formamide, Methanol, and Formic Acid. *J. Phys. Chem.*, 82(5):592–601, 1978.
- [398] S. R. Cox and D. E. Williams. Representation of the Molecular Electrostatic Potential by a Net Atomic Charge Model. *J. Comput. Chem.*, 2(3):304–323, 1981.
- [399] U. C. Singh and P. A. Kollman. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.*, 5(2):129–145, 1984.
- [400] Björn Bieniek. *Ultra Thin ZnO on Metal Substrates: An Ab Initio Study*. PhD thesis, Technische Universität Berlin (TU Berlin), 2016.

- 
- [401] M. Mantina, A. C. Chamberlin, R. Valero, C. J. Cramer, and D. G. Truhlar. Consistent van der Waals Radii for the Whole Main Group. *J. Phys. Chem. A*, 113(19):5806–5812, 2009.
- [402] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.
- [403] H. Joshi. *Functional Domain Motions and Processivity in Bacterial Hyaluronate Lyase: A Molecular Dynamics Study*. Universal Publishers, 2010.
- [404] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A Well-behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.*, 97(40):10269–10280, 1993.
- [405] F.-Y. Dupradeau, A. Pigache, T. Zaffran, C. Savineau, R. Lelong, N. Grivel, D. Lelong, W. Rosanski, and P. Cieplak. The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.*, 12:7821–7839, 2010.
- [406] E. Sigfridsson and U. Ryde. Comparison of Methods for Deriving Atomic Charges from the Electrostatic Potential and Moments. *J. Comput. Chem.*, 19(4):377–395, 1998.
- [407] P. M. Morse. Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels. *Phys. Rev.*, 34:57–64, 1929.
- [408] S. Jakobsen, K. Kristensen, and F. Jensen. Electrostatic Potential of Insulin: Exploring the Limitations of Density Functional Theory and Force Field Methods. *J. Chem. Theory Comput.*, 9(9):3978–3985, 2013.
- [409] R. S. Mulliken. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.*, 23(10):1833–1840, 1955.
- [410] E. Sedghamiz, B. Nagy, and F. Jensen. Probing the Importance of Charge Flux in Force Field Modeling. *J. Chem. Theory Comput.*, 13(8):3715–3721, 2017.





# Acknowledgments

First and foremost, I owe my deepest gratitude to PD Dr. Carsten Baldauf who is not only an exceptional group leader but also a humble and kind person essential for guiding my way through sometimes difficult times that come with research and writing a thesis.

I would particularly like to thank Prof. Dr. Thomas Rizzo who not only offered his knowledge and input as the thesis co-director but also provided me with the opportunity to get hands-on experience in the field of conformer-selective UV-IR spectroscopy within his group at EPFL.

I want to thank Prof. Dr. Matthias Scheffler for giving me the opportunity to do interesting research at the Theory Department of the Fritz Haber Institute.

I thank Prof. Dr. Volker Blum who as my first group leader introduced me to the world of *ab initio* biomolecular simulations.

I am very grateful to the internal and external examiners of the thesis jury, Prof. Dr. Michele Ceriotti, Prof. Dr. Gert von Helden, and Dr. Lubomír Rulíšek for their time and interest in the scientific topics presented.

I am grateful to the Max Planck-EPFL Center for Molecular Nanoscience and Technology – and in particular to Prof. Dr. Klaus Kern and Dr. Klaus Kuhnke – for the funding of the joint research project.

I wish to thank Steffen Kangowski for always being very helpful with all hard- and software related questions of mine.

The co-workers at the Theory Department of the Fritz Haber Institute have always provided a pleasant working atmosphere. There are too many people to thank including (and please forgive me if I forgot anyone) Dmitrii, Adriana, Franziska, Johanna, Arvid (putting the “pro” in programming), Björn, Wael, Oliver, Franz, Igor, Hanna, Julia, Birgit, Maria, Luca, Majid, Hagen-Henrik, Honghui, Christian, Sebastian, and Lydia. A special thanks also to Mateusz Marianski with whom it was a pleasure to work with and to have an after-work-beer on the balcony.

## Acknowledgments

---

During my stay at EPFL, I had the pleasure to get to know wonderful people who gave me a fantastic time there, in particular Lucy, Val, Maarten, Ana, and Jörn.

I would like to express my heartfelt appreciation to Chiara Masellis who I had the most awesome time with when taking spectra until 3am for almost two weeks straight. It was certainly THE most fun and yet surprisingly productive time imaginable. Inoltre, ti ho detto che ti avrei scritto una frase in italiano. L'unica cosa di cui mi pento è non aver inserito la Coca-Cola nella macchina per misurare la fenilalanina.

Mein tiefster Dank gilt meinen Eltern, die mir und meiner Familie immer und in jeglicher Form zur Seite standen. Danke für alles von Herzen!

Es ist unmöglich für mich die Dankbarkeit und Liebe in Worte zu fassen, die ich für meine Kinder Mattis und Elinor und für meine Frau Nancy empfinde. Ich danke Dir, Nancy, Du hast viel auf Dich genommen und ohne Dich hätte ich es nicht durch diese schwierige Zeit geschafft.

## Curriculum Vitae

---

### Markus Schneider

#### Diploma (M.Sc.) in Physics

Date of Birth: October 3rd, 1984

Place of Birth: Riesa, Germany

Current Address:  
Springbornstr. 226  
12487 Berlin  
Germany

E-Mail: markus.schneider@fhi-berlin.mpg.de

### Current Position:

---

since October 2014 **Fritz Haber Institute of the Max Planck Society, Berlin, Germany**

PhD student

- Research interests:
  - o Structural changes of biomolecules in the gas-phase in presence of cations
  - o Energy benchmarks of biomolecular systems
  - o Machine learning of force field parameters from *ab initio* calculations

### Education

---

Oct 2006 – Sept 2011 **Humboldt-Universität zu Berlin, Germany**

Main Studies of Physics

- Degree: Diploma (M.Sc.) with grade 1,1 (A)

Oct 2010 – Sept 2011 **Deutsches Elektronen-Synchrotron (DESY), Zeuthen, Germany**

**Diploma Thesis** with Grade 1,0 (A)

“Effects of Hadronization, Multiple-Parton-Interactions, Pile-Up and Unfolding Corrections in Multi-jet Events at the ATLAS Detector”

Oct 2005 – Sept 2006 **Universidad Autónoma de Madrid, Spain**

Continuation of Physics Studies

Oct 2003 – Sept 2006 **Technische Universität Dresden, Germany**

Studies (basic courses) of Physics

Lausanne, 16 Juillet 2018



### Publications

- M. Schneider and C. Baldauf. Relative Energetics of Acetyl-Histidine Protomers with and without  $\text{Zn}^{2+}$  and a Benchmark of Energy Methods. *to be submitted*, 2018.
- M. Schneider and C. Baldauf. Force Field Parameterization Using Regularized Linear Regression. *to be submitted*, 2018.
- M. Schneider, C. Masellis, T. Rizzo, C. Baldauf. Kinetically Trapped Liquid-State Conformers of a Sodiated Model Peptide Observed in the Gas Phase. *J. Phys. Chem. A*, 121(36):6838–6844, 2017.
- M. Marianski, A. Supady, T. Ingram, M. Schneider, C. Baldauf. Assessing the Accuracy of Across-the-Scale Methods for Predicting Carbohydrate Conformational Energies on the Example of Glucose and  $\alpha$ -Maltose. *J. Chem. Theory Comput.*, 12(12):6157–6168, 2016.
- M. Ropo, M. Schneider, C. Baldauf, V. Blum. First-Principles Data Set of 45,892 Isolated and Cation-Coordinated Conformers of 20 Proteinogenic Amino Acids. *Sci. Data*, 3:160009, 2016.
- J. Klyne, A. Bouchet, S. Ishiushi, M. Fujii, M. Schneider, C. Baldauf, O. Dopfer. Probing Chirality Recognition of Protonated Glutamic Acid Dimers by Gas-Phase Vibrational Spectroscopy and First-Principles Simulations. *to be submitted*, 2018.
- The ATLAS Collaboration. Measurement of Multi-Jet Cross Sections in Proton-Proton Collisions at a 7 TeV Center-of-Mass Energy. *Eur. Phys. J. C – Particles and Fields*, 71(11):1763, 2011.

## Conference Contributions – Posters

- “Theoretical Conformational Search of Heptapeptide-Cation Systems and Comparison with Experiment”. Conference on Molecular Nanostructures 2017, Ascona, Switzerland, 2017.
- “Improving Force-Field and First-Principles Driven Conformational Sampling Using Compressive Sensing”. Summer School on Interfaces and Energy, Göttingen, Germany, 2016.
- “Force-Field and First-Principles Driven Conformational Sampling of Microsolvated Histidine-Cation Systems”. Summer School of the Max-Planck-EPFL Center for Molecular Nanoscience & Technology, Ringberg, Germany, 2015.
- “Force-Field and First-Principles Driven Conformational Sampling of Microsolvated Histidine-Cation Systems”. Psi-k Conference, San Sebastián, Spain, 2015.
- “Force-Field and First-Principles Driven Conformational Sampling of Microsolvated Histidine-Cation Systems”. Bunsentagung, Bochum, Germany, 2015.
- “Histidine-Cation Interaction and Microsolvation from First Principles”, White Nights of Materials Science, St. Petersburg, Russia, 2014.
- “Histidine-Cation Interaction and Microsolvation from First Principles”, Molecular Nanosystems Workshop, Ascona, Switzerland, 2014.
- “Histidine-Cation Interaction and Microsolvation from First Principles”, DPG Spring Meeting, Berlin, Germany, 2014.

## Workshop Contributions – Invited Talks

- “Peptide-Cation Systems: Conformational Search, Benchmark Evaluation, and Force Field Parameter Adjustment Using Regularized Linear Regression”. Workshop of the FHI Theory Department 2017, Liepe, Germany, 2017.
- “High-level Quantum Chemistry Methods and Benchmark Datasets for Molecules”. Hands-on Workshop and Humboldt-Kolleg: Density-Functional Theory and Beyond - Basic Principles and Modern Insights, Isfahan University of Technology, Isfahan, Iran, 2016.