Metal Cations in Protein Force Fields: From Data Set Creation and Benchmarks to Polarizable Force Field Implementation and Adjustment

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

Submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

Xiaojuan Hu

from Hangzhou, China.

March, 2023

This work was performed between May 2018 and March 2023 under the supervision of PD Dr. Baldauf at Fritz-Haber-Institut der Max-Planck-Gesellschaft.

1st reviewer: PD Dr. Carsten Baldauf

2nd reviewer: Prof. Dr. Bettina Keller

Date of defense: 03.07.2023

My doctoral degree thesis entitled "Metal Cations in Protein Force Fields: From Data Set Creation, Force Field Benchmarks, to Polarizable Force Field Implementation and Adjustment" has been prepared by myself and it is based on my own work. The work by others has been specifically acknowledged in the text. The thesis is submitted to the Department of Biology, Chemistry, Pharmacy of Freie Universität Berlin to obtain the academic degree Doctor rerum naturalium (Dr. rer. nat.) and has not been submitted for any other degree.

Berlin, March 2023

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor Carsten Baldauf for giving me the opportunity to perform research in his group at the Fritz Haber Institute of the Max Planck Society. His valuable suggestions, guidance, discussions, and encouragement walked me through all the stages of the research. I would like to extend my appreciation to Prof. Dr. Bettina Keller for kindly agreeing to be my co-supervisor and reviewer of my thesis. I'm truly grateful to Dr. Mariana Rossi for all the group meetings and group outings in SABIA group that have enriched my project and my life. She is always available to discuss and help me with any doubts and problems. Furthermore, I would like to thank Dr. Markus Schneider, the first developer of FFAFFURR. His excellent work provided a solid basis for my project. I would like to thank Dmitrii Maksimov, who helped me to start my research and helped me a lot with conformational sampling. We work closely with Kazi Shudipto Amin on force field development and molecular dynamics. I want to thank him for all the fruitful discussions and ideas. For the ontology part of this thesis, I am very grateful to Dr. Maja-Olivia Lenz-Himmer. She gave me great help with ontology development. She is always willing to discuss and help. Markus Scheidgen helped us upload our huge dataset to NOMAD. I want to thank him for all his help and quick responses related to NOMAD. I'm grateful to everyone from the SABIA group: Haiyuan Wang, Alaa Akkoush, Karen Fidanyan, Yair Litman, Nathaniel Raimbault, Marcin Krynski. They all helped me with my project. They are not only good colleagues, but also good friends. I would like to thank all of my collaborators for their fruitful discussions, valuable contributions and great insights into the project.

I want to thank the administrative staff at the FHI: Julia Pach, Hanna Krauter, Annika Scior, Steffen Kangoswki and Rayya Douedari. I appreciate the computer infrastructure support from Max Planck Society (in MPCDF), impressive seminars hold at FHI, and access to articles guaranteed by ArXiv and Sci-hub. I'm grateful to everyone in the FHI NOMAD laboratory for the shared coffee breaks and group outings. I'm grateful to my Chinese friends in Berlin: Yuanyuan Zhou, Fangfang Huo, Sheng Bi, Zhenkun Yuan, and many more. They helped me a lot and made me feel at home. I am also grateful to every nice person I met in Berlin who made me love this beautiful city more and more.

Last but not the least, I would like to thank my beloved family for supporting me in every decision I make.

Abstract

Metal cations are essential to life. About one-third of all proteins require metal cofactors to accurately fold or to function. Computer simulations using empirical parameters and classical molecular mechanics models (force fields) are the standard tool to investigate proteins' structural dynamics and functions *in silico*. Despite many successes, the accuracy of force fields is limited when cations are involved. The focus of this thesis is the development of tools and strategies to create system-specific force field parameters to accurately describe cation-protein interactions. The accuracy of a force field mainly relies on (i) the parameters derived from increasingly large quantum chemistry or experimental data and (ii) the physics behind the energy formula.

The first part of this thesis presents a large and comprehensive quantum chemistry data set on a consistent computational footing that can be used for force field parameterization and benchmarking. The data set covers dipeptides of the 20 proteinogenic amino acids with different possible side chain protonation states, 3 divalent cations (Ca^{2+}, Mg^{2+}, and Ba^{2+}), and a wide relative energy range. Crucial properties related to force field development, such as partial charges, interaction energies, etc., are also provided. To make the data available, the data set was uploaded to the NOMAD repository and its data structure was formalized in an ontology.

Besides a proper data basis for parameterization, the physics covered by the terms of the additive force field formulation model impacts its applicability. The second part of this thesis benchmarks three popular non-polarizable force fields and the polarizable Drude model against a quantum chemistry data set. After some adjustments, the Drude model was found to reproduce the reference interaction energy substantially better than the non-polarizable force fields, which showed the importance of explicitly addressing polarization effects. Tweaking of the Drude model involved Boltzmann-weighted fitting to optimize Thole factors and Lennard-Jones parameters. The obtained parameters were validated by (i) their ability to reproduce reference interaction energies and (ii) molecular dynamics simulations of the N-lobe of calmodulin. This work facilitates the improvement of polarizable force fields for cation-protein interactions by quantum chemistry-driven parameterization combined with molecular dynamics simulations in the condensed phase.

While the Drude model exhibits its potential simulating cation-protein interactions, it lacks description of charge transfer effects, which are significant between cation and protein. The CTPOL model extends the classical force field formulation by charge transfer (CT) and polarization (POL). Since the CTPOL model is not readily available in any of the popular molecular-dynamics packages, it was implemented in OpenMM. Furthermore, an open-source parameterization tool, called FFAFFURR, was implemented that enables the (system specific) parameterization of OPLS-AA and CTPOL models. Following the method established in the previous part, the performance of FFAFFURR was evaluated by its ability to reproduce quantum chemistry energies and molecular dynamics simulations of the zinc finger protein.

In conclusion, this thesis steps towards the development of next-generation force fields to accurately describe cation-protein interactions by providing (i) reference data, (ii) a force field model that includes charge transfer and polarization, and (iii) a freely-available parameterization tool.

Kurzzusammenfassung

Metallkationen sind für das Leben unerlässlich. Etwa ein Drittel aller Proteine benötigen Metall-Cofaktoren, um sich korrekt zu falten oder zu funktionieren. Computersimulationen unter Verwendung empirischer Parameter und klassischer Molekülmechanik-Modelle (Kraftfelder) sind ein Standardwerkzeug zur Untersuchung der strukturellen Dynamik und Funktionen von Proteinen *in silico*. Trotz vieler Erfolge ist die Genauigkeit der Kraftfelder begrenzt, wenn Kationen beteiligt sind. Der Schwerpunkt dieser Arbeit liegt auf der Entwicklung von Werkzeugen und Strategien zur Erstellung systemspezifischer Kraftfeldparameter zur genaueren Beschreibung von Kationen-Protein-Wechselwirkungen. Die Genauigkeit eines Kraftfelds hängt hauptsächlich von (i) den Parametern ab, die aus immer größeren quantenchemischen oder experimentellen Daten abgeleitet werden, und (ii) der Physik hinter der Kraftfeld-Formel.

Im ersten Teil dieser Arbeit wird ein großer und umfassender quantenchemischer Datensatz auf einer konsistenten rechnerischen Grundlage vorgestellt, der für die Parametrisierung und das Benchmarking von Kraftfeldern verwendet werden kann. Der Datensatz umfasst Dipeptide der 20 proteinogenen Aminosäuren mit verschiedenen möglichen Seitenketten-Protonierungszuständen, 3 zweiwertige Kationen (Ca^{2+}, Mg^{2+} und Ba^{2+}) und einen breiten relativen Energiebereich. Wichtige Eigenschaften für die Entwicklung von Kraftfeldern, wie Wechselwirkungsenergien, Partialladungen usw., werden ebenfalls bereitgestellt. Um die Daten verfügbar zu machen, wurde der Datensatz in das NOMAD-Repository hochgeladen und seine Datenstruktur wurde in einer Ontologie formalisiert.

Neben einer geeigneten Datenbasis für die Parametrisierung beeinflusst die Physik, die von den Termen des additiven Kraftfeld-Modells abgedeckt wird, dessen Anwendbarkeit. Der zweite Teil dieser Arbeit vergleicht drei populäre nichtpolarisierbare Kraftfelder und das polarisierbare Drude-Modell mit einem Datensatz aus der Quantenchemie. Nach einigen Anpassungen stellte sich heraus, dass das Drude-Modell die Referenzwechselwirkungsenergie wesentlich besser reproduziert als die nichtpolarisierbaren Kraftfelder, was zeigt, wie wichtig es ist, Polarisationseffekte explizit zu berücksichtigen. Die Anpassung des Drude-Modells umfasste eine Boltzmanngewichtete Optimierung der Thole-Faktoren und Lennard-Jones-Parameter. Die erhaltenen Parameter wurden validiert durch (i) ihre Fähigkeit, Referenzwechselwirkungsenergien zu reproduzieren und (ii) Molekulardynamik-Simulationen des Calmodulin-N-Lobe. Diese Arbeit demonstriert die Verbesserung polarisierbarer Kraftfelder für Kationen-Protein-Wechselwirkungen durch quantenchemisch gesteuerte Parametrisierung in Kombination mit Molekulardynamiksimulationen in der kondensierten Phase.

Während das Drude-Modell sein Potenzial bei der Simulation von Kation - Protein - Wechselwirkungen zeigt, fehlt ihm die Beschreibung von Ladungstransfereffekten, die zwischen Kation und Protein von Bedeutung sind. Das CTPOL-Modell erweitert die klassische Kraftfeldformulierung um den Ladungstransfer (CT) und die Polarisation (POL). Da das CTPOL-Modell in keinem der gängigen Molekulardynamik-Pakete verfügbar ist, wurde es in OpenMM implementiert. Außerdem wurde ein Open-Source-Parametrisierungswerkzeug namens FFAFFURR implementiert, welches

die (systemspezifische) Parametrisierung von OPLS-AA- und CTPOL-Modellen ermöglicht. In Anlehnung an die im vorangegangenen Teil etablierte Methode wurde die Leistung von FFAF-FURR anhand seiner Fähigkeit, quantenchemische Energien und Molekulardynamiksimulationen des Zinkfingerproteins zu reproduzieren, bewertet.

Zusammenfassend lässt sich sagen, dass diese Arbeit einen Schritt in Richtung der Entwicklung von Kraftfeldern der nächsten Generation zur genauen Beschreibung von Kationen-Protein-Wechselwirkungen darstellt, indem sie (i) Referenzdaten, (ii) ein Kraftfeldmodell, das Ladungstransfer und Polarisation einschließt, und (iii) ein frei verfügbares Parametrisierungswerkzeug bereitstellt.

Contents

Abbreviation List

- AAMI: amino acid meta-info
- AD: Alzheimer's disease
- AIM: atoms-in-molecule
- API: application programming interface
- BOA: Born-Oppenheimer approximation
- CC: coupled cluster
- CI: configuration interaction
- CN: coordination number
- CTPOL: charge transfer and polarization
- DFA: density functional approximation
- DFT: density functional theory
- ECC: empirical continuum correction
- EMMO: European materials modelling ontology
- ESP: electrostatic potential
- FC: fluctuating charge
- FF: force field
- FFAFFURR: framework for adjusting force fields using regularized regression
- FHI-aims: Fritz Haber Institute ab initio molecular simulations
- FQ: fluctuating charge
- GA: genetic algorithm
- GGA: generalized gradient approximation
- HF: Hartree-Fock
- KS: Kohn-Sham
- LASSO: least absolute shrinkage and selection operator
- LDA: local density approximation
- LJ: Lennard-Jones
- MAE: mean absolute error
- MBPT: many-body perturbation theory
- MC: Monte Carlo
- MD: molecular dynamics
- ME: maximum error
- MM: molecular mechanics
- MP2: second-order Møller-Plesset
- MT: metallothionein
- PDB: protein data bank
- PES: potential energy surface
- PSO: particle swarm optimization
- QM: quantum mechanical
- REMD: replica exchange molecular dynamics
- TS: Tkatchenko–Scheffler
- vdW: van der Waals
- xc: exchange-correlation

List of Publications

Paper I "Better force fields start with better data: A data set of cation dipeptide interactions"

X. Hu, M. O. Lenz-Himmer, C. Baldauf.

*Sci. Data **9**, 1-14 (2022)*

Paper II "Benchmarking polarizable and non-polarizable force fields for Ca^{2+}–peptides against a comprehensive QM dataset"

K. S. Amin, **X. Hu**, D. R. Salahub, C. Baldauf, C. Lim and S. Noskov.

*J. Chem. Phys. **153**, 144102 (2020)*

Paper III "System-specific parameter optimization for non-polarizable and polarizable force fields"

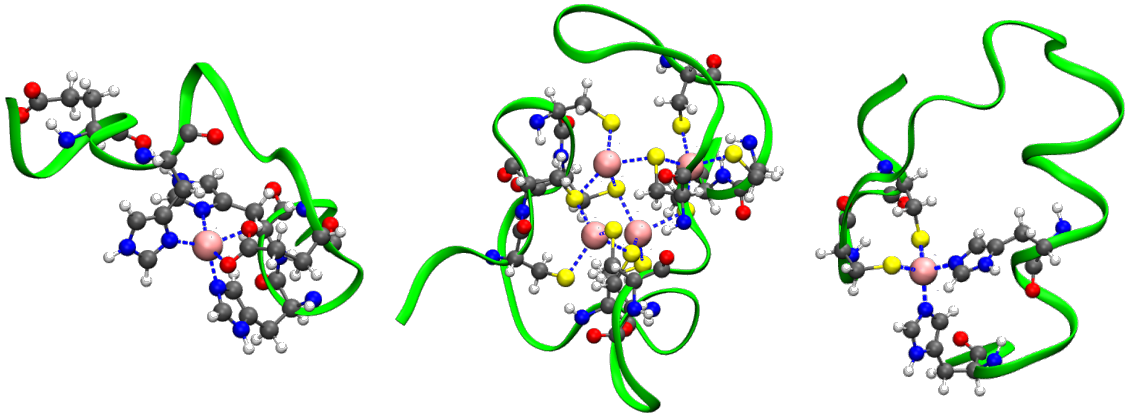
X. Hu, K. S. Amin, M. Schneider, C. Lim, D. Salahub and C. Baldauf.

arXiv preprint arXiv:2303.12775 (2023).

Chapter 1

Introduction

Metal cations are essential to life. About one-third of the proteins in Protein Data Bank (PDB) contain metal cations,[1] which typically play a crucial role in shaping the three-dimensional structure of proteins and peptides, and therefore affect important properties, e.g. binding sites, catalytic properties, and biological functions. A systematic bioinformatics survey reveals that among 1,371 different enzymes, 47% require metal cations to maintain their three-dimensional structure, and 41% are known to rely on metal cations in their catalytic centers.[2] As an abundant cation in the human body, zinc cations are required for the functional centers of more than 200 enzymes, for example carbonic anhydrase, alkaline phosphatase, and glycerol phosphate dehydrogenase.[3–5] Exemplary, the Aβ sequences, group of Metallothioneins (MTs), and Zinc finger protein are discussed here. The progressive neurodegenerative disease Alzheimer's disease (AD) is associated with the formation of senile plaques, which dominantly consist of aggregated Aβ_{40}/Aβ_{42} in the brain.[6,7] Numerous studies[8–11] have reported that Zn^{2+}, Cu^{2+}, and Fe^{3+} may act as the seeding factor of Aβ plaques and the existence of Zn^{2+} enhances Aβ aggregation. Histidine (His), Glutamate (Glu), and Aspartic Acid (Asp) are the potential binding sites of Zn^{2+} in the Aβ sequence. Figure 1 (a) shows the structure of the Aβ(1-16)-Zn^{2+} complex. There are three His residues and one Glu residue as binding centers that coordinate Zn^{2+}. MTs were discovered in 1957 and were identified as a family of low-molecular-weight, cysteine-rich, and metal-rich proteins present in all living organisms.[12,13] MTs play a role in protecting cells and tissues from heavy metal toxicity, maintaining the homeostasis of intracellular free Zn^{2+}, and controlling neuronal growth. There is growing evidence that MTs play important roles in various human tumors, drug resistance, and neurodegenerative diseases such as AD.[14] In mammals, MTs are divided into four groups according to their encoding genes: MT-1, MT-2, MT-3, and MT-4. Due to the high cysteine (Cys) content (30%), MTs bind a variety of trace metals including zinc, cadmium, mercury, platinum, and silver. MT-1 and MT-2 mainly bind Zn^{2+}, and Cd^{2+} to a lesser extent. MT-3 binds Zn^{2+} and Cu^{2+} equally well.[15] The structure of the β domain of human MT-2 is shown in Figure 1 (b). Eleven deprotonated cysteines (Cys) are binding to 4 Zn^{2+} in the center of the β domain of human MT-2. Zinc finger proteins are one of the most abundant protein groups. They can interact with



(a) (b) (c)

Figure 1: Structures of (a) of Aβ(1-16)-Zn^{2+} complex (PDB ID: 1ZE9), (b) β domain of human MT-1 (PDB ID: 1MHU), and (c) zinc finger protein (PDB ID: 1ZNF).

DNA, RNA, and other proteins and thus participate in many cellular processes, including DNA recognition, signal transduction, DNA repair, and so on.[16] The structure of zinc finger proteins is maintained by the zinc center coordinating cysteine and histidine. The structure of one of the zinc finger proteins is shown in Figure 1 (c).

The detailed analysis of structure, dynamics, and function of the metal coordination architecture within metalloproteins is an important addition to the understanding of metalloprotein functions. Besides excellent experimental studies, computer simulations are playing important roles in chemical research and life science. Computational chemistry provides insights from the electronic level to even beyond the molecular level that are difficult or impossible to observe experimentally, thus it complements experiments and provides further insight into underlying mechanisms. It is desirable to reach such a detailed and fundamental theoretical understanding also of cation-peptide interaction systems. Numerous computational studies investigated metalloproteins. For example, Tamames *et al.*[17] investigated the structural characteristics of Zn coordination spheres by a thorough analysis on a data set of 994 proteins from the Protein Data Bank, complemented with DFT calculations at the B3LYP/SDD level. Baldauf *et al.*[18] studied the underlying nature of metal cations altering peptide structures. Zhou *et al.*[19] found that the Ca^{2+} binding site in the blood protein von Willebrand Factor (VWF) regulates force-triggered unfolding for cleavage by classical force-probe molecular dynamics (MD) simulations.

Classical MD simulations at the atomic scale are widely used to study the conformational dynamics of biomolecular systems. They reveal mechanisms that are difficult to observe experimentally on small spatial and temporal scales.[20] MD simulations have gained many successes ranging from protein folding and aggregation[21,22] to transmembrane protein dynamics.[23] However, MD simulations fail to reproduce or predict experimental results in some cases. These inadequacies stem from the statistical errors due to the finite length of simulations and the systematic errors caused by inaccurate models employed.[20] Over the past 20 years, the development of hardware (the computational power of central processing units (CPUs) doubles every 18 months according

to Moore's law[24]), optimized software[25–28] and the enhanced sampling methods[29] have significantly reduced the statistical errors so that the systematic errors are detectable and alleviated. Empirically parameterized force fields are typically employed in MD simulations because of their speed advantage, which allows for larger system size, and longer simulation time scale compared to quantum chemistry-based simulations. In spite of the many successes that have been made with force fields in the simulation of bio-systems, the accuracy of force fields is far from ideal, especially when it comes to the interactions of ionic species.[30–33]

The applicability of force fields depends on several factors, one of them is the accuracy and scope of the parameterization data employed.[20] No matter if the force field parameters were fitted to computational or experimental reference data, the systems under investigation are typically different from the training data. For example, the Lennard-Jones (LJ) parameters in OPLS were derived from the vaporization calorimetry of pure organic liquids such as pyridine, benzene, or tetrahydrofuran, while these parameters were later applied to the analysis of sugars, oligopeptides, or proteins.[34,35] The types of reference molecules that are employed in some of the classical force fields are listed in Table 1. The assumption of transferability means that similar substructures of different systems can be represented by the same set of parameters. Furthermore, many parameters of existing force fields are derived based on comparably small reference data sets, which creates further uncertainty about the reliability of force fields. Parameters derived from a large and high-quality data set, where molecules are as chemically similar as possible to the target system and that cover a large relative energy range, may lead to more reliable force field results. Experimental data are often limited, especially for metal complexes, and typically describe low energy conformers, while transition states are lacking. Consequently, electronic structure calculations are often being used for force field parameterization due to their advantages of good accuracy with affordable computational cost.[36] In the case of proteins, their individual building blocks have been investigated in many studies. For example, Rezac *et al.*[37] created a QM data set containing several smaller peptides and medium-sized macrocycles that can be of potential use for force field parameterization and assessed the performance of popular QM methods. Kishor *et al.*[38] investigated conformations, energetics, and ionization potential of 20 amino acids with density functional theory. These and similar studies have deepened our understanding of the fundamental structural basic of peptides and proteins. However, the level of theory and the sampling methods applied are highly diverse in these studies. For metal-cation containing systems, a sufficient amount of experimental or computational data is lacking. Furthermore, the data is often not available in a usable way.

The classical FF energy is composed as a sum of the so-called bonded and non-bonded interactions. The details of FF methods are explained in section 2.1.6. Non-bonded interactions are crucial for simulating the behavior of metalloprotein systems. The interactions between metal ions and surrounding atoms are modeled as the sum of van der Waals (vdW) interactions and electrostatic interactions. The strategies of deriving LJ parameters and partial charges in classical force fields are listed in Table 2. The vdW interactions are typically described by the popular

Table 1: List of reference molecules of fixed-charge force fields.

Family	Reference molecules type	Version	Year
GROMOS	amino acids	26C1 [39]	1982
	amino acids, nucleic acids, lipids	53A5/53A6 [40]	2004
	small molecules	ATB1.0 [41]	2011
CHARMM	amino acids	CHARMM22 [42]	1992
	small molecules	CGenFF [43]	2010
AMBER	amino acids, nucleic acids	ff99 [44]	1999
	small molecules	GAFF [45]	2004
OPLS	small molecules, amino acids	OPLS-AA [34]	1996
	carbohydrates	OPLS-AA [46]	1997

12-6 LJ functional form in classical force fields. LJ parameters are usually obtained by fitting to experimental properties. In recent years, approaches for deriving LJ parameters from quantum chemistry calculations, for example employing the atoms-in-molecule (AIM) method,[47] have been proposed. Electrostatic interactions are described by the Coulomb potentials. Partial charges are usually derived by two strategies: (i) fitting to experimental data, e.g. hydration free enthalpies[40] (GROMOS 53A5/53A6), or (ii) derivation from quantum chemistry calculations (AMBER GAFF[45] and CHARMM22[42]). These two strategies can also be used in combination (OPLS after 2005).[48] Since the first strategy is time-consuming and requires extensive testing using Monte Carlo (MC) or MD simulations to reproduce the target experimental properties, partial charge extraction from quantum chemistry calculations tends to be preferred in the development of modern force fields.[49] Using nonbonded interactions to simulate ionic interactions is able to reproduce the monovalent situation closely. However, for divalent ion systems with higher electronegativity, the quality of the classical force field decreases. One approach to overcome the limitations of classical force fields in describing metalloprotein systems is to refine the classical force field parameters. Empirical Continuum Correction (ECC)[50–52] force fields take electronic polarization into account implicitly in a mean-field way by scaling the charge of cations and residues. The force-matching method[53,54] improves on the classical force fields by adjusting parameters to reproduce *ab initio* forces. Some approaches[55,56] refine the LJ parameters, or add a $1/r^4$ term to the standard 12-6 LJ potential, resulting in a 12–6–4 LJ-type model to account for charge-induced dipole interactions. Notably, LJ parameters and partial charges are intrinsically correlated and thus must be treated jointly.[57] All of these refinements have been successful to some extent. However, they are still limited in describing the high diversity of electrostatic environments in metalloproteins.

The limitations of classical force fields on simulating metalloproteins also stem from the underlying assumption: the atomic charges are fixed. However, effects like polarization and charge transfer, which proved to be very important for ionic systems, are ignored.[64–66] Polarizable force fields explicitly include polarization effects and allow the simulation of charge delocalization effects in response to environmental changes. There are basically three classes of polarizable models: the fluctuating charge (FQ) model, the induced dipole model, and the Drude oscillator model. The FQ model[67,68] allows redistribution of atomic charges to equalize electronegativity in response to environmental changes, while maintaining overall charge conservation. In this way, charge transfer

Table 2: List of parameterization strategies used for electrostatic interactions and LJ interactions in the fixed-charge force fields.

Force field version	Partial charge parameterization	LJ parameterization
GROMOS 53A5/53A6 [40]	fitting liquid and hydration properties	fitting liquid and hydration properties
GROMOS ATB1.0 [41]	B3LYP/6-31G* in implicit water with ESP	taken from 53A6
CHARMM22 [42]	HF/6-31G* in vacuum with ESP, scaling by 1.16	fitting liquid properties
CHARMM CGenFF [43]	trained BCI	fitting liquid properties
AMBER ff99 [44]	HF/6-31G* in vacuum with RESP	taken from OPLS-AA
AMBER GAFF [45]	HF/6-31G* in vacuum with RESP or AM1-BCC	taken from ff99
OPLS-AA [34]	fitting liquid and hydration properties, gas-phase geometries, complexation energies	fitting liquid and hydration properties gas-phase geometries, complexation energies
OPLS2.0 [48]	CM1A-BCC	taken from OPLS-AA

ESP: charges fitted to electrostatic potential. [58]
RESP: charges derived by a restrained electrostatic potential fitting procedure. [59]
BCI: charges assigned using a bond-charge increment (BCI) scheme. [43]
AM1-BCC: AM1 atomic charges [60] with bond charge corrections (BCCs) added. [61,62]
CM1A-BCC: a combination of the semiempirical CM1A [63] charge and BCC. [48]

is simulated dynamically. The FQ models include: CHARMM-FQ, [69] OPLS-AA-FQ, [70] ABEEMsp (atom-bond electronegativity equalization model with s- and p-bonds), [71] etc. The FQ model is one of the simplest polarizable models. It is orders of magnitude faster than quantum chemistry calculations, while still creating reliable atomic charges for a set of compounds. [72] However, FQ models usually overestimate dipole moments and failed to simulate out-of-plane polarization effects, which are very important for the simulation of many functional groups such as aromatic rings. [69,73] Although researchers have attempted to include out-of-plane effects by adding virtual charge sites, it has been proven to be inefficient due to the challenges of scaling simulations of large systems. [71]

The induced dipole model calculates the electrostatic energy for each site based on its induced dipole and the electrostatic field at that site. [65] Well-known induced dipole models include the AMOEBA (atomic multipole optimized energetics for biomolecular simulation) force field [74] and the SIBFA (sum of interactions between fragments ab initio) model. [75] The induced dipole can be represented by different strategies, for example, whether to consider higher multipole moments or whether to consider higher order inductions (e.g. the induced quadrupole). In principle, including higher order terms gives better accuracy, while the difficulty of parameterization and computational cost increase. The Drude oscillator model simulates induced polarization effects by attaching a particle carrying a charge to polarizable (heavy) atom via a harmonic spring. Beyond only including charge transfer or polarization effects, there are models that include both of them. For example, the charge transfer and polarization (CTPOL) model [64,76] incorporates charge transfer and local polarization effects into the framework of classical force field. The charge transfer between ligand atoms and metal ions is significant if strong charge donors such as thiolate groups coordinating to metal ions are present in metalloproteins. [77] However, the inclusion of charge transfer into a classical force field reduces the amount of charge on the metal connecting atoms and ions, thereby weakening their charge/dipole-charge interactions. This can be compensated by introducing local

polarization energies of metal ions and ligand atoms.

Numerous studies have shown that polarizable force fields yield better accuracy than classical force fields, especially for systems containing divalent ions. For example, Jing *et al.*[78] used the AMOEBA model to predict the Ca^{2+} and Mg^{2+} selectivity in protein, whereas the classical force field model AMBER failed even after parameterization. Yu *et al.*[79] derived Drude parameters for a large set of monovalent and divalent cations and demonstrated the good performance of the Drude model for simulating ionic solvation in aqueous solutions. Ponder *et al.*[80] showed that although further refinement is necessary to reproduce solvation free energies of drug-like molecules, the AMOEBA force field is especially successful in predicting protein-ligand poses in comparison to classical force fields. Polarizable models have been shown to predict the water dimer energy with similar accuracy as quantum chemistry calculations.[81] However, models including polarization do not always perform better than classical models, depending on the quality of the parameters. Furthermore, polarizable models have received limited validation, which implies that there is still room for improvement in the polarizable models. Reparameterization may be required when applied to different systems, and polarizable models have more parameters and thus more elaborate parameterization schemes.

Overall, we see two directions to improve the accuracy of simulations of metalloprotein systems:

- Include more physics in the force field formula, for example charge transfer and polarization effects.
- Provide a sufficiently-accurate and available electronic-structure data set that covers a wide range of conformational space to parameterize the force field.

Based on these points, this thesis starts with creating a uniform and comprehensive quantum chemistry data set of amino-methylated and acetylated (capped) dipeptides of the 20 proteinogenic amino acids with various possible protonation states and their interactions with selected divalent cations.[82] The data set covers a wide range of relative energies and properties relevant to force field development. To make the data set accessible even to experts from other fields, an ontological representation of the data set is provided. The details of this work are shown in section 4.1. In the second work as shown in section 4.2,[83] Ca^{2+}-dipeptide systems from the data set were employed to benchmark the polarizable Drude FF and three widely used classical FFs, namely, OPLS-AA, AMBER, and CHARMM (C36). In this work, we demonstrated improved accuracy by adjusting parameters of Drude FF and by the explicit account of charge-transfer and polarization effects (CTPOL) of the simulation of cation-dipeptide systems. Finally, since the parameterization is always time-consuming and labor-intensive, section 4.3 shows a developed open-source parameterization tool that enables the parameterization of OPLS-AA and CTPOL models.

This doctoral thesis is organized as follows: Chapter 2 contains the theoretical background and the methodology applied in this thesis. Chapter 3 summarizes the main results of publications. The publications and contributions of each author are outlined in Chapter 4. Finally, the conclusion and outlook of this thesis are presented in Chapter 5.

Chapter 2

Theoretical Background

2.1 Methods to Calculate Potential Energy Surface

One of the major problems in computational chemistry is to select an appropriate level of theory for a given problem.[84] In our case, it is the choice of an appropriate method to sample and describe the characteristics of the ion-dipeptide systems. This section will try to evaluate potential energy surface (PES) with different theoretical models and levels of theory.

2.1.1 The Many-Body Hamiltonian

Most coupled electron-nucleus systems in the chemistry field can be described by a many-body non-relativistic Hamiltonian. The Hamilton operator $\hat{H}$ consists of five terms:[85]

$$\hat{H} = \hat{T}_{\mathrm{n}} + \hat{V}_{\mathrm{n-n}} + \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{n-e}} + \hat{V}_{\mathrm{e-e}}, \tag{2.1}$$

where $\hat{T}_{\mathrm{n}}$ represents the nuclear kinetic energy operator, $\hat{T}_{\mathrm{e}}$ represents the electronic kinetic energy operators, $\hat{V}_{\mathrm{n-n}}$, $\hat{V}_{\mathrm{n-e}}$, and $\hat{V}_{\mathrm{e-e}}$ represent the spin-independent Coulombic interaction between nucleus-nucleus, electron-nucleus, and electron-electron, respectively. With the use of natural units,[86] *i.e.*

$$\begin{aligned} \hbar &= 1, \\ m_e &= 1, \\ |e| &= 1, \\ 4\pi\epsilon_0 &= 1, \end{aligned} \tag{2.2}$$

these operators can be written as

$$\begin{aligned}
\hat{T}_{\mathrm{n}} &= \sum_{k=1}^{M} \frac{(-i\hbar\nabla_{\vec{R}_k})^2}{2M_k} = -\sum_{k=1}^{M} \frac{\nabla^2_{\vec{R}_k}}{2M_k} \\
\hat{T}_{\mathrm{e}} &= \sum_{j=1}^{N} \frac{(-i\hbar\nabla_{\vec{r}_j})^2}{2m_e} = -\sum_{j=1}^{N} \frac{\nabla^2_{\vec{r}_j}}{2} \\
\hat{V}_{\mathrm{n-n}} &= \frac{1}{2}\sum_{k_1\neq k_2}^{M} \frac{1}{4\pi\epsilon_0}\frac{Z_{k_1}Z_{k_2}e^2}{|\vec{R}_{k_1}-\vec{R}_{k_2}|} = \frac{1}{2}\sum_{k_1\neq k_2}^{M} \frac{Z_{k_1}Z_{k_2}}{|\vec{R}_{k_1}-\vec{R}_{k_2}|} \\
\hat{V}_{\mathrm{n-e}} &= -\sum_{k=1}^{M}\sum_{j=1}^{N} \frac{1}{4\pi\epsilon_0}\frac{Z_k e^2}{|\vec{R}_k-\vec{r}_j|} = -\sum_{k=1}^{M}\sum_{j=1}^{N} \frac{Z_k}{|\vec{R}_k-\vec{r}_j|} \\
\hat{V}_{\mathrm{e-e}} &= \frac{1}{2}\sum_{j_1\neq j_2}^{N} \frac{1}{4\pi\epsilon_0}\frac{e^2}{|\vec{r}_{j_1}-\vec{r}_{j_2}|} = \frac{1}{2}\sum_{j_1\neq j_2}^{N} \frac{1}{|\vec{r}_{j_1}-\vec{r}_{j_2}|},
\end{aligned} \tag{2.3}$$

where $\vec{R}_k$, M_k, Z_k, and k are the position vector, mass, charge, and index for the M nuclear, and $\vec{r}_j$, m_e, $-e$ and j are the position vector, mass, charge and index for the N electrons. Solving this Hamiltonian in a non-relativistic and time-independent quantum-mechanical framework means solving the time-independent Schrödinger equation:

$$\hat{H}\Psi = E\Psi, \tag{2.4}$$

where E denotes the total energy and Ψ represents the many-body wave function of the system. While the Schrödinger equation has $3M+3N$ degrees of freedom and the solution is not separable in its variables. Thus, the exact solutions are only available for a few limited cases and approximations have to be made to deal with it.

2.1.2 Born-Oppenheimer Approximation

The standard first step in solving the Schrödinger equation is to partially decouple the electron from the nuclear motion. This is achieved via the Born-Oppenheimer approximation.[87]

The Born-Oppenheimer approximation relies on the fact that electrons are thousand times lighter than a nucleus. Thus, electrons move much faster than nuclei, which means that the electrons adapt to the movement of the nuclei instantaneously and, therefore, it is assumed the movement of the nuclei cannot induce any electronic excitation. For this reason, the Born-Oppenheimer approximation is also called adiabatic approximation. The many-body wave function Ψ can then be separated into the nuclear wave function Ψ_{n} and the electron wave function Ψ_{e}:

$$\Psi(\vec{R}_1,\ldots,\vec{R}_M,\vec{r}_1,\ldots,\vec{r}_N) = \Psi_{\mathrm{n}}(\vec{R}_1,\ldots,\vec{R}_M)\Psi_{\mathrm{e}}(\vec{R}_1,\ldots,\vec{R}_M,\vec{r}_1,\ldots,\vec{r}_N). \tag{2.5}$$

This allows the electronic part to be solved with the electron wave function depending only parametrically on the nuclear coordinates.

The Schrödinger equation can be written as

$$\left[\left(\hat{T}_{\mathrm{n}}+\hat{V}_{\mathrm{n-n}}\right) + \underbrace{\left(\hat{T}_{\mathrm{e}}+\hat{V}_{\mathrm{e-e}}+\hat{V}_{\mathrm{n-e}}\right)}_{\hat{H}_{\mathrm{e}}}\right]\Psi(\vec{R}_1,\ldots,\vec{R}_M,\vec{r}_1,\ldots,\vec{r}_N) = E\Psi(\vec{R}_1,\ldots,\vec{R}_M,\vec{r}_1,\ldots,\vec{r}_N), \tag{2.6}$$

in which the Hamilton operator is divided into two parts: the first part only depends on the nuclear coordinates $\vec{R}_1, \ldots, \vec{R}_M$, while the latter part, which is defined as electronic Hamiltonian $\hat{H}_\mathrm{e}$, also depends on the electronic coordinates $\vec{r}_1, \ldots, \vec{r}_N$. Insert Equation 2.5 into Equation 2.6, and completely neglect the nuclear kinetic energy because of the adiabatic approximation, which means assuming $\hat{T}_\mathrm{n}\Psi(\vec{R}_1, \ldots, \vec{R}_M, \vec{r}_1, \ldots, \vec{r}_N)$ can be neglected, the total energy of the system is given as:

$$\begin{aligned} E &= \hat{V}_{\mathrm{n-n}} + E_\mathrm{e}(\vec{R}_1, \ldots, \vec{R}_M) \\ &= \frac{1}{2} \sum_{k_1 \neq k_2}^{M} \frac{Z_{k_1} Z_{k_2}}{|\vec{R}_{k_1} - \vec{R}_{k_2}|} + E_\mathrm{e}(\vec{R}_1, \ldots, \vec{R}_M). \end{aligned} \tag{2.7}$$

The electronic energy $E_\mathrm{e}(\vec{R}_1, \ldots, \vec{R}_M)$ can be obtained by solving the electronic Schrödinger equation:

$$\hat{H}_\mathrm{e}\Psi_\mathrm{e}(\vec{R}_1, \ldots, \vec{R}_M, \vec{r}_1, \ldots, \vec{r}_N) = E_\mathrm{e}(\vec{R}_1, \ldots, \vec{R}_M)\Psi_\mathrm{e}(\vec{R}_1, \ldots, \vec{R}_M, \vec{r}_1, \ldots, \vec{r}_N). \tag{2.8}$$

Solving Equation 2.8 is a very difficult computational task. The challenge lies in the huge number and the quantum nature of the electrons. For this reason, even if the nuclear motion is ignored, an efficient handling of the electron problem is necessary. In the following sections, various approaches to obtain approximate solutions to Equation 2.8 will be described.

2.1.3 Hartree-Fock methods

For a given Hamiltonian $\hat{H}$, the ground state energy E_0 is the minimum expectation value that can be achieved for any normalized wave function, *i.e.*

$$E_0 = \min_{\Psi} \langle \Psi | \hat{H} | \Psi \rangle, \tag{2.9}$$

where

$$\langle \Psi | \hat{H} | \Psi \rangle = \int \Psi^* \hat{H} \Psi d\vec{r}^N. \tag{2.10}$$

The Dirac notation, or *bra-ket* notation, is used to simplify the notation. One of the oldest methods to find the ground state wave function is based on this and is called the variational principle.[85] In this approach, a set of trial normalized wave functions that depend on several parameters is considered, and the expectation value of the energy is minimized in order to find the ground state wave function and the corresponding energy.

An important simplification towards solving Equation 2.8 is the introduction of the independent-particle model, which assumes that the motion of one electron is independent of the motion of all other electrons. This means that the interactions between electrons are approximated, either by ignoring all but the most important one or by taking the average interaction. While only the latter has acceptable accuracy and is known as Hartree-Fock (HF) theory.[88] In the Hartree-Fock model, the wave function is described as the product of the single-particle wave functions:

$$\Psi_\mathrm{e}^{\mathrm{HF}}(\vec{r}_1, \ldots, \vec{r}_N) = \psi_1(\vec{r}_1)\psi_2(\vec{r}_2)\ldots\psi_n(\vec{r}_n), \tag{2.11}$$

where the explicit parametric dependence on $\vec{R}_1, \ldots, \vec{R}_M$ has been omitted for simplicity. ψ_j represents a single particle wave function or electron orbital. However, the Hartree Ansatz in Equation 2.11 does not fulfill the Pauli principle,[89] which states that two electrons can not have all quantum numbers equal, by not taking the indistinguishability of electrons into account. In other words, the total electronic wave function must be antisymmetric. This antisymmetry requirement of the electronic wave function can be achieved by a Slater determinant:

$$\Psi_{\mathrm{e}}^{\mathrm{HF}}(\vec{r}_1, \ldots, \vec{r}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\vec{r_1}) & \psi_2(\vec{r_1}) & \ldots & \psi_N(\vec{r_1}) \\ \psi_1(\vec{r_2}) & \psi_2(\vec{r_2}) & \ldots & \psi_N(\vec{r_2}) \\ \ldots & \ldots & \ldots & \ldots \\ \psi_1(\vec{r_N}) & \psi_2(\vec{r_N}) & \ldots & \psi_N(\vec{r_N}) \end{vmatrix}. \tag{2.12}$$

The spin dependencies have been ignored throughout. The method of finding the electronic ground state by the variational principle using a Slater determinant as the ansatz of the wave function is known as the Hartree-Fock method. The ground state energy $E_{\mathrm{e}}^{\mathrm{HF}}$ can be written as:

$$E_{\mathrm{e}}^{\mathrm{HF}} = \langle \Psi_0^{\mathrm{H}} | \hat{H}_{\mathrm{e}} | \Psi_0^{\mathrm{H}} \rangle = \sum_{i=1}^{N} H_i + \frac{1}{2} \sum_{i,j=1}^{N} (J_{ij} - K_{ij}), \tag{2.13}$$

where

$$\begin{aligned} H_i &= \int \psi_i^*(\vec{r}) \left[-\frac{1}{2}\nabla^2 - \sum_{k=1}^{M} \frac{Z_k}{|\vec{R}_k - \vec{r}|} \right] \psi_i(\vec{r}) d\vec{r} \\ J_{ij} &= \iint \psi_i^*(\vec{r}) \psi_j^*(\vec{r}') \frac{1}{|\vec{r} - \vec{r}'|} \psi_i(\vec{r}) \psi_j(\vec{r}') d\vec{r} d\vec{r}' \\ K_{ij} &= \iint \psi_i^*(\vec{r}) \psi_j^*(\vec{r}') \frac{1}{|\vec{r} - \vec{r}'|} \psi_j(\vec{r}) \psi_i(\vec{r}') d\vec{r} d\vec{r}'. \end{aligned} \tag{2.14}$$

J_{ij} represents the Coulomb integral and K_{ij} represents the exchange integral. It should be noted that $J_{ij} \geq K_{ij} \geq 0$ and $J_{ii} = K_{ii}$. The Coulomb "self-interaction" J_{ii} is canceled by the corresponding exchange term K_{ii}. Thus, the HF method is said to be self-interaction free. With the definition of Hartree energy E_{Hartree} and exchange energy E_{x}, Equation 2.13 can be written as:

$$E_{\mathrm{e}}^{\mathrm{HF}} = \sum_{i=1}^{N} H_i + E_{\mathrm{Hartree}} + E_{\mathrm{x}}, \tag{2.15}$$

where

$$\begin{aligned} E_{\mathrm{Hartree}} &= \frac{1}{2} \sum_{i,j=1}^{N} J_{ij} \\ E_{\mathrm{x}} &= -\frac{1}{2} \sum_{i,j=1}^{N} K_{ij}. \end{aligned} \tag{2.16}$$

The next step is to find a set of orbitals that minimize the energy, with the constraint that all electron orbitals ψ_i are orthonormal. The Fock operator $\hat{F}$ is an effective one-electron energy operator. It is associated with the variation of the total energy and is given by:

$$\hat{F} = -\frac{1}{2}\nabla^2 - \sum_{k=1}^{M} \frac{Z_k}{|\vec{R}_k - \vec{r}|} + \hat{j} - \hat{k}, \tag{2.17}$$

where the Coulomb operator $\hat{j}$ and the exchange operator $\hat{k}$ are given as:

$$\begin{aligned}\hat{j}(\vec{r})f(\vec{r}) &= \sum_{i=1}^{N}\int \psi_i^*(\vec{r}')\psi_i(\vec{r})\frac{1}{|\vec{r}-\vec{r}'|}f(\vec{r})d\vec{r}' \\ \hat{k}(\vec{r})f(\vec{r}) &= \sum_{i=1}^{N}\int \psi_i^*(\vec{r}')f(\vec{r})\frac{1}{|\vec{r}-\vec{r}'|}\psi_i(\vec{r})d\vec{r}',\end{aligned} \tag{2.18}$$

in which $f(\vec{r})$ is an arbitrary function. With the Fock operator, the Hartree-Fock differential equations can be written as:

$$\hat{F}\psi_i(\vec{r}) = \sum_{j=1}^{N}\epsilon_{ij}\psi_j(\vec{r}), \tag{2.19}$$

where ϵ_{ij} are the Lagrange multipliers and is given by

$$\epsilon_{ij} = \sigma_{ij}\epsilon_j. \tag{2.20}$$

Thus, the Hartree-Fock equations become

$$\hat{F}\psi_i(\vec{r}) = \sum_{j=1}^{N}\epsilon_i\psi_i(\vec{r}). \tag{2.21}$$

ϵ_i represents orbital energies of the single non-interacting electron orbitals.

Solving the Hartree-Fock equations is an eigenvalue problem. However, the Fock operator depends on all orbitals (via the Coulomb and exchange operators). Thus, an iterative method must be employed to solve the problem. With guessed initial orbitals ψ_i, the Fock operator can be generated, which leads to new orbitals by solving the Hartree-Fock equations in Equation 2.21. The new orbitals lead to an updated Fock operator. This process repeats until convergence.

It is clear that the total energy cannot be exact because the electron–electron repulsion is only considered in an average way and consequently neglects the correlation between electrons. The HF model is a kind of branching point in which either more approximations are involved, leading to semi-empirical methods, or more determinants are added, thereby leading stepwise to the exact solution of the electronic Schrödinger equation. The latter one is the so-called "post-Hartree-Fock" techniques, which will be briefly discussed in the next section.

2.1.4 Post-Hartree-Fock Methods

As mentioned before, the HF model fails to capture the electron correlation. Thus, it can not well describe systems that have strongly correlated electrons in the context of this work, in particular, *e.g.* hydrogen-bonded and systems involving biomolecules. The correlation energy, $E_{\mathrm{e}}^{\mathrm{corr}}$, is defined as

$$E_{\mathrm{e}}^{\mathrm{corr}} = E_{\mathrm{e}} - E_{\mathrm{e}}^{\mathrm{HF}}, \tag{2.22}$$

where E_{e} represents the exact electronic energy in Born-Oppenheimer approximation and $E_{\mathrm{e}}^{\mathrm{HF}}$ represents the Hartree-Fock energy given in Equation 2.13. Various methods have attempted to capture the correlation energy based on the HF model.

The HF wave function is a determinant of the low energy orbitals or "occupied orbitals". The starting point for improving HF results is obviously to include more Slater determinants. The new

series of determinants may be constructed by replacing one or more orbitals that are occupied in the HF determinant with "unoccupied orbitals". Depending on how many "occupied orbitals" are replaced by "unoccupied orbitals", the determinants are often referred to as Singles (S), Doubles (D), Triples (T), Quadruples (Q), *etc.* The multi-determinant trial wave function can be written as a linear combination of the HF wave function, Ψ_{HF}, and other determinants, Ψ_i,

$$\Psi = a_0\Psi_{\mathrm{HF}} + \sum_{i=1} a_i\Psi_i. \tag{2.23}$$

Three main methods are used to describe electron correlation: Configuration Interaction (CI), Many-Body Perturbation Theory (MBPT), and Coupled Cluster (CC).

CI[90] is based on the variational principle. The trial wave function is a linear combination of determinants with expansion coefficients. The expansion coefficients are optimized to make the energy a minimum. If all the possible electronic configurations are considered in the wave function, this method is called full-configuration interaction (full-CI). Finding all expansion coefficients using the variational principle is computationally extremely demanding, and truncating the expansion by including excitations of only several electrons leads to size-consistency problems, i.e., the energy of two non-interacting molecules is not twice the energy of one of them calculated at the same level of approximation.

MBPT defines a Hamilton operator that consists of two parts, an unperturbed (H_0) and a perturbation (H'). The perturbation operator H' is assumed to be smaller than H_0 and can be added as a correction by employing an independent-particle approximation. To apply perturbation theory, the unperturbed Hamiltonian must be selected. The most common choice is to sum up the Fock operators, resulting in Møller–Plesset (MP) perturbation theory.[91] The second-order Møller-Plesset (MP2) is a simple alternative to the full-CI method. It is the lowest non-vanishing correction term to HF. For systems with a few hundred basis functions, the cost of MP2 can be similar or lower than the cost of HF. MP2 typically can grasp 80–90 % of the correlation energy. However, it does not follow the variational principle, which means that it is possible to find energy lower than the exact energy given by the Born-Oppenheimer approximation. Furthermore, it overestimates the correlation energy in systems containing anions, strongly electronegative atoms, or transition metals, and it cannot be employed to describe metallic systems.

CC is not based on the variational principle but guarantees size consistency.[92] It starts with the definition of excitation operator $\hat{T}$

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \dots. \tag{2.24}$$

The i-th excitation operator T_i acting on the HF wave function Slater determinant generates all excited Slater determinants

$$\hat{T}_1\Psi_{\mathrm{e}}^{\mathrm{HF}} = \sum_{i}^{occ.}\sum_{\alpha}^{unocc.} t_i^{\alpha}\Psi_i^{\alpha}, \tag{2.25}$$

where Ψ_i^{α} represents the Slater determinant in which the "occupied orbital" i is replaced by "unoccupied orbital" α, and t_i^{α} is the corresponding coefficient. The wave function ansatz of

CC is expressed as:

$$\Psi_{\mathrm{CC}} = e^{\hat{T}}\Psi_{\mathrm{e}}^{\mathrm{HF}} = (1 + \hat{T} + \frac{\hat{T}^2}{2!} + \frac{\hat{T}^3}{3!} + \dots)\Psi_{\mathrm{e}}^{\mathrm{HF}}. \tag{2.26}$$

The most commonly used expansion is CCSD(T)[93] where $\hat{T}$ is truncated at the Singles (S) and Doubles (D) excitation levels, i.e., $\hat{T} = \hat{T}_1 + \hat{T}_2$ is solved, and triple excitation, $\hat{T}_3$, is added using Møller-Plesset perturbation theory. It provides excellent accuracy for non-covalent systems.[94] Thus, it is often referred to as the "gold standard of quantum chemistry". However, it is formally scaled as $O(N^7)$, where N represents the size of the system, which results in extremely expensive computations. In order to reduce the computational costs while maintaining the accuracy, many efforts have been made, such as the proposed domain-based local pair natural orbital (DLPNO-)CCSD(T)[95] approximation, which shows a near-linear scaling behavior with system size N. CCSD(T) is often used as benchmark to validate approximations of lower-level, such as density functional theory (DFT) methods. DFT will be presented in the next section as the main electronic structure method in this thesis.

2.1.5 Density Functional Theory

The solution of the electronic Schrödinger equation (Equation 2.8) is the wave function Ψ_{e}, which has $4N$ variables for a system containing N electrons, $3N$ spatial and N spin coordinates. The complexity of a wave function increases exponentially with the number of electrons, making it very difficult to describe. Density Functional Theory (DFT) is an electronic-structure method that replaces the complex N-electron wave function Ψ_{e} with the electron density $\rho(\vec{r})$, which only depends on 3 spatial coordinates.

Hohenberg and Kohn developed the theoretical footing of DFT and proved that it is possible to calculate all the properties of systems with electron densities through their two well-known theorems:[96]

1. The external potential is uniquely specified for a given ground state electron density $\rho(\vec{r})$.

2. The electron density that gives the energy minimum is the exact ground state density ρ_0.

The proofs of these two theorems can be found in Reference[97]. Hohenberg and Kohn theorems do not provide any practical use for obtaining the ground state energy or density since the functional for the electronic energy is not provided.

The Kohn-Sham (KS) scheme provides a practical way to connect electron density with ground-state energy. The use of a set of auxiliary orbitals to calculate the electron kinetic energy was proposed by Kohn and Sham in 1965, which laid the foundation for the success of modern DFT methods.[98] The electronic density can then be calculated as a sum of single-particle KS orbitals:

$$\rho(\vec{r}) = \sum_{i}^{N} \psi_i^*(\vec{r})\psi_i(\vec{r}). \tag{2.27}$$

The total energy $E_{\mathrm{e}}[\rho(\vec{r})]$ can be rewritten as a functional of the electron density $\rho(\vec{r})$

$$\begin{aligned} E_{\mathrm{e}}[\rho(\vec{r})] &= T[\rho(\vec{r})] + E_{\mathrm{H}}[\rho(\vec{r})] + E_{\mathrm{ext}}[\rho(\vec{r})] + E_{\mathrm{xc}}[\rho(\vec{r})] \\ &= -\frac{1}{2}\sum_{i}^{N}\langle\psi_i^*(\vec{r})|\nabla^2|\psi_i(\vec{r})\rangle + \frac{1}{2}\iint\frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}d\vec{r}' + \int V_{\mathrm{ext}}(\vec{r})\rho(\vec{r})d\vec{r} + E_{\mathrm{xc}}[\rho(\vec{r})], \end{aligned} \tag{2.28}$$

where $T[\rho(\vec{r})]$ represents the kinetic energy, $E_{\mathrm{H}}[\rho(\vec{r})]$ represents the Coulomb interaction energy or Hartree term, $E_{\mathrm{ext}}[\rho(\vec{r})]$ is the interaction energy caused by the external potential $V_{\mathrm{ext}}(\vec{r})$, and all the many-body complexities are addressed by the exchange-correlation functional $E_{\mathrm{xc}}[\rho(\vec{r})]$, which is still unknown. Equation 2.28 is possible to be minimized with respect to the electron density under the constraint $\int d\vec{r}\rho(\vec{r}) = n$ with the variational principle. Similar to the Hartree-Fock theory, KS equations can be reduced to a system of single-particle equations,

$$\left(-\frac{1}{2}\nabla^2 + v_{\mathrm{H}}(\vec{r}) + v_{\mathrm{ext}}(\vec{r}) + v_{\mathrm{xc}}(\vec{r})\right)\psi_i(\vec{r}) = \epsilon_i\psi_i(\vec{r}), \tag{2.29}$$

$$\begin{aligned} v_{\mathrm{H}}(\vec{r}) &= \frac{\partial E_{\mathrm{H}}[\rho]}{\partial\rho(\vec{r})} \\ v_{\mathrm{xc}}(\vec{r}) &= \frac{\partial E_{\mathrm{xc}}[\rho]}{\partial\rho(\vec{r})}, \end{aligned} \tag{2.30}$$

where $\psi_i(\vec{r})$ is the KS spatial orbital, and ϵ_i is the orbital energy. The total energy then can be expressed as a function of the eigenvalues:

$$E_{\mathrm{e}}[\rho(\vec{r})] = \sum_{i}^{N}\epsilon_i - \frac{1}{2}\int\rho(\vec{r})v_{\mathrm{H}}(\vec{r})d\vec{r} - \int\rho(\vec{r})v_{\mathrm{xc}}(\vec{r})d\vec{r} + E_{\mathrm{xc}}[\rho(\vec{r})], \tag{2.31}$$

where the double-counting terms are subtracted from the sum of the eigenvalues. KS equations have to be solved self-consistently. Starting with a trial electron density, Equation 2.29 is solved and thereby a new set of KS orbitals that yield an updated electron density. This procedure is repeated until the total energy is minimized self-consistently.

DFT is a true *ab initio* technique if the exact expression of the exchange-correlation (xc) functional would be known. The main deficiency of DFT is that the exact solution can not be obtained. Many approximations have been made, which result in different density-functional approximations (DFA). These approximations define the accuracy of DFA and are explained in the following.

Local Density Approximation (LDA)

The Local Density Approximation (LDA) treats the electron density as a uniform electron gas. The exchange-correlation energy functional $E_{\mathrm{xc}}^{\mathrm{LDA}}$ in LDA is given as:

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int\epsilon_{\mathrm{xc}}[\rho(\vec{r})]\rho(\vec{r})d\vec{r}, \tag{2.32}$$

where the the exchange-correlation energy of each particle $\epsilon_{\mathrm{xc}}[\rho(\vec{r})]$ is the energy of the uniform electron gas. $\epsilon_{\mathrm{xc}}[\rho(\vec{r})]$ can be divided into two parts, exchange and correlation contributions,

$$\epsilon_{\mathrm{xc}}[\rho(\vec{r})] = \epsilon_{\mathrm{x}}[\rho(\vec{r})] + \epsilon_{\mathrm{c}}[\rho(\vec{r})], \tag{2.33}$$

which leads to

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = E_{\mathrm{x}}^{\mathrm{LDA}}[\rho] + E_{\mathrm{c}}^{\mathrm{LDA}}[\rho]. \tag{2.34}$$

The exchange energy part has an analytical form, which can be written as:

$$E_{\mathrm{x}}^{\mathrm{LDA}}[\rho] = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int \rho^{4/3}(\vec{r})d\vec{r}. \tag{2.35}$$

The analytical form of the correlation energy is not known, but there are approximations, e.g. the PZ-LDA[99] and PW-LDA[100] approximations, both obtained from quantum Monte Carlo calculations[101] and VWN-LDA approximation.[102] The LDA method is a good approximation for systems where the electron density changes slowly, such as bulk metals. However, LDA fails for systems where the electron density can not be treated as uniform. For example, molecular systems where dispersion interactions are important.

The Generalized Gradient Approximation (GGA)

To improve the LDA, a non-uniform electron gas must be considered. One step in this direction is the Generalized Gradient Approximation (GGA), which includes the gradients of the electron density as a variable in the xc functional. The xc functional is given as:

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[\rho] = \int \rho(\vec{r})\epsilon_{\mathrm{xc}}[\rho(\vec{r})]f_{\mathrm{xc}}[\rho(\vec{r}), \nabla\rho(\vec{r})]d\vec{r}, \tag{2.36}$$

where ϵ_{xc} is the functional of the homogeneous electron gas, and f_{xc} is the factor enhancement, which varies in different GGA parameterization. One of the most widely used GGA xc functional is the Perdew-Burke-Ernzerhof (PBE) functional.[103] PBE is a non-empirical functional, which means that all parameters are basic constants and there are no empirical parameters. In most cases, GGA functionals show improvements over LDA in several properties, e.g. binding energies, atomic energies, and energy barriers. Although GGA methods yield good results when analyzing the structure of molecules, they are known to underestimate the binding energy of systems that have weak interactions like hydrogen bonds.

Van der Waals Correction Schemes in Density-Functional Theory

Despite the many successes of DFT, DFAs in standard use are not well constructed to describe long-range electron correlation effects. In particular, it can not properly describe long-range dispersion effects or vdW interactions.[104] While a good treatment of weak vdW interactions is crucial for an accurate energetic description of biomolecules.[105–107] The physical nature of the attractive vdW interactions arises from the long-range part of the correlation of electronic density fluctuations.

One straightforward way to account for vdW interactions is to add empirical or semi-empirical corrections to the DFA energy. Thus, the total energy is given as:

$$E_{\mathrm{tot}} = E_{\mathrm{DFA}} + E_{\mathrm{vdW}}, \tag{2.37}$$

where E_{DFA} represents the total energy yielded in DFA, and E_{vdW} is the dispersion correction. Since the dominant term of vdW dispersion interaction is the instantaneous dipole-induced dipole

interaction, the vdW energy is proportional to the well-known $1/R^6$ potential:

$$E_{\text{vdW}} = -\frac{1}{2}\sum_{A\neq B}\frac{C_{6,AB}}{R_{AB}^6} f_{damp}(R_{AB}), \tag{2.38}$$

where A and B represent two different atoms, R_{AB} is the interatomic distance between two atoms, $C_{6,AB}$ is the heteronuclear dispersion coefficient, and f_{damp} is the damping function. f_{damp} is employed to couple the short-range, which is mainly described by DFA, and long-range contributions of the electron correlation. Thus it fulfills:

$$\begin{aligned} f_{damp}(R_{AB}) &\underset{R_{AB}\to 0}{\to} 0 \\ \text{and } f_{damp}(R_{AB}) &\underset{R_{AB}\to \infty}{\to} 1. \end{aligned} \tag{2.39}$$

Based on the definitions of $C_{6,AB}$ and f_{damp}, several approaches exist.

In the Casimir-Polder formula,[108] the expression of the coefficient $C_{6,AB}$ is:

$$C_{6,AB} = \frac{3}{\pi}\int_0^\infty \alpha_A(i\omega)\alpha_B(i\omega)d\omega, \tag{2.40}$$

where $\alpha_A(i\omega)$ is the average dynamic polarizability, which can be measured experimentally, and ω is the excitation frequency. Keeping only the leading term of the Padé series,[109] the polarizability can be approximated as:

$$\alpha_A(\omega) = \frac{\alpha_A^0}{1-(\omega/\omega_A)^2}, \tag{2.41}$$

where α_A^0 denotes the static polarizability and ω_A is effective excitation frequency. Inserting Equation 2.41 into Equation 2.40 and solving the integral analytically, the $C_{6,AB}$ can be written as:

$$C_{6,AB} = \frac{3}{2}\alpha_A^0\alpha_B^0\frac{\omega_A\omega_B}{\omega_A+\omega_B}. \tag{2.42}$$

The effective excitation frequency ω_A of atom A for homonuclear $C_{6,AA}$ can be written as:

$$\omega_A = \frac{4}{3}\frac{C_{6,AA}}{(\alpha_A^0)^2}. \tag{2.43}$$

Inserting Equation 2.43 into Equation 2.42 yields:

$$C_{6,AB} = \frac{2C_{6,AA}C_{6,BB}}{\left(\frac{\alpha_B^0}{\alpha_A^0}C_{6,AA} + \frac{\alpha_A^0}{\alpha_B^0}C_{6,BB}\right)}. \tag{2.44}$$

In this way, the $C_{6,AB}$ can be calculated with the free-atom parameters α_A^0 and $C_{6,AA}$.[110]

Most schemes[111–115] employ fixed empirical C_6 coefficients for each atom. However, the actual dispersion coefficients depend on the molecular environment around the atoms.[116] The use of fixed C_6 coefficients leads to errors in dispersion interaction estimates. The popular Grimme scheme[117] introduces DFA-specific global scaling parameter to reduce the functional dependence of the scheme.

A popular scheme proposed by Becke and Johnson in 2005[116,118] uses non-empirical dispersion C_6 coefficients. The C_6 coefficients in this scheme are obtained from the exchange-hole dipole moment. Thus the KS-orbitals and the density of a system need to be provided. The damping function is still empirical.

The parameter-free pairwise Tkatchenko-Scheffler (TS) van der Waals scheme (vdW^{TS})[119] incorporates ideas from both strategies above and is used in this thesis. Using the definition of effective atomic volume, the formulation can be adjusted to be environment-dependent for an atom in a molecule. With the atomic Hirshfeld partitioning scheme,[120,121] the ratio of the effective atomic volume (V_A) referenced to the free atom (V_A^{free}) *in vacuo* $\frac{V_A}{V_A^{\text{free}}}$ is given by:

$$\frac{V_A}{V_A^{\text{free}}} = \frac{\int r^3 \omega_A(\vec{r}) \rho(\vec{r}) d\vec{r}}{\int r^3 \rho_A^{\text{free}}(\vec{r}) d\vec{r}}, \tag{2.45}$$

where r is the distance from the nucleus of atom A and a point, $\rho(\vec{r})$ is the total electron density, $\rho_A^{free}(\vec{r})$ is the electron density of the free atom A, and $\omega_A(\vec{r})$ is the Hirshfeld atomic partitioning weight, which is given by:

$$\omega_A(\vec{r}) = \frac{\rho_A^{\text{free}}(\vec{r})}{\sum_B^{\text{all atoms}} \rho_B^{\text{free}}(\vec{r})}. \tag{2.46}$$

With the definition of $\frac{V_A}{V_A^{free}}$, the effective quantities are defined as:

$$C_{6,AA} = \left(\frac{V_A}{V_A^{\text{free}}}\right)^2 C_{6,AA}^{\text{free}} \tag{2.47}$$

$$R_A^0 = \left(\frac{V_A}{V_A^{\text{free}}}\right)^{1/3} R_A^{0,\text{free}}, \tag{2.48}$$

where R denotes the vdW radius. In this way, the C_6 coefficients related to the electron density still remain partly empirical due to the use of Hirshfeld partitioning.

In the TS scheme, the damping function is a Fermi-type function:

$$f_{damp}^{AB}\left(R_{AB}, R_{AB}^0(\rho(\vec{r}))\right) = \frac{1}{1 + \exp\left[-d\left(\frac{R_{AB}}{s_R R_{AB}^0(\rho(\vec{r}))} - 1\right)\right]}, \tag{2.49}$$

where R_{AB} denotes the distance between atoms A and B, R_{AB}^0 is the sum of the effective vdW radii associated with atoms A and B (Equation 2.48), d is a free parameter affecting the steepness of the damping and has been set to $d = 20$, and s_R is a free empirical scaling coefficient that adjusts the extent of the vdW correction for a given xc functional and is obtained by fitting to the S22 data set.[122] The TS scheme has been tested on a database that contains 1225 intermolecular C6 pairs and showed a mean absolute error of 5.5% compared to experimental results regardless of the employed xc functional.[119]

Basis set

The single-particle orbitals ψ_i can be expanded by a set of basis functions:

$$\psi_i = \sum_n c_{ni} \phi_n(\boldsymbol{r}) \tag{2.50}$$

All the DFT calculations in this thesis were performed with the all-electron code FHI-aims,[123] which uses localized numeric atom-centered orbital (NAO) basis set. The minimal NAO basis is composed of the core and functions of spherically symmetric free atoms and is exact for free atoms. The shapes of the orbitals close to the nuclei are well described also for bonded atoms using the

minimal NAO basis. The basis functions in FHI-aims are ordered as *tiers* according to the amount of improvement that the basis functions yield to the total energy of dimers. The basis functions in *tier1* bring the largest improvement, while the basis functions in *tier4* bring the smallest but still noticeable improvements. This thesis employs two different sets of numerical defaults for atomic species: *light* and *tight* settings. In the *light* settings, *tier1* is used and the integration grids are not so dense. The *tier2* basis sets are utilized in the *tight* settings, and the integration grids are more dense. The *tight* settings yield converged results and can be used for production calculations.

2.1.6 Force Fields

Unlike the methods described above, force field methods, also known as molecular mechanics (MM) methods, do not treat the electrons of the system explicitly. They use atoms or groups of atoms as the "building blocks". The energy of a system is written as a parametric function of the nuclear coordinates. The dynamics of atoms are described with classical mechanics, i.e. Newton's second law. Force field methods are especially useful when *ab initio* methods are unfeasible because of the high computational cost and limited time scale, for example, conformational sampling and MD simulations of proteins. Several classical and polarizable force field models are introduced in this section.

Classical Force Fields

The classical (bio)molecular force field energy, E_{FF}, can be written as a sum of energy terms, each of them corresponding to the energy required to distort a molecule in a specific way:

$$E_{\mathrm{FF}} = E_{\mathrm{bonds}} + E_{\mathrm{angles}} + E_{\mathrm{tors}} + E_{\mathrm{improper}} + E_{\mathrm{vdW}} + E_{\mathrm{ele}} \tag{2.51}$$

where E_{bonds} is the energy required for stretching a bond between two atoms, E_{angles} corresponding to the energy for bending an angle, E_{tors} represents the energy for rotation around a bond, E_{vdW} and E_{ele} describe the nonbonded interactions between atoms.

We see that a force field is a combination of individual bonded and non-bonded terms that need a set of parameters and atomic coordinates as input. The parameters in force fields are derived by fitting to experimental data and/or higher-level quantum chemistry data. With the assumption of transferability, i.e. structurally similar atoms in different molecules may share parameter values, parameters are usually fitted for small molecules and subsequently applied in larger ones.

Conventional (classical) force fields that are commonly applied to describe biomolecular systems include CHARMM36[124] (Chemistry at Harvard Macromolecular Mechanics 36), AMBER-99[125] (Assisted Model Building with Energy Refinement 99), OPLS-AA[126] (Optimized Potentials for Liquid Simulations - All-Atom), and so on. They all share a similar functional form. Here, we take

OPLS-AA as an example, the bonded terms are the following:

$$E_{\text{bonds}} = \sum_{bonds}^{1-2atoms} \frac{1}{2} K_{ij}^{r} \left(r_{ij} - r_{ij}^{0}\right)^2 \tag{2.52}$$

$$E_{\text{angles}} = \sum_{angles}^{1-3atoms} \frac{1}{2} K_{ij}^{\theta} \left(\theta_{ij} - \theta_{ij}^{0}\right)^2 \tag{2.53}$$

$$E_{\text{tors}} = \sum_{dihedrals,n}^{1-4atoms} V_{n}^{ij} \left(1 + cos\left(n\phi_{ij} - \phi_{ij}^{0}\right)\right) \tag{2.54}$$

$$E_{\text{improper}} = \sum_{improper}^{1-4atoms} V_{2imp}^{ij} \left(1 + cos\left(2\phi_{ij} - \phi_{ij}^{0}\right)\right) \tag{2.55}$$

As shown in Equation 2.52, E_{bonds} is in the form of a harmonic oscillator, with the potential being quadratic in the displacement of the bond length r_{ij} from the reference length r_{ij}^{0}. K_{ij}^{r} is the force constant for the $i-j$ bond. Similarly, i and j in Equation 2.53 are atoms separated by two bonds. K_{ij}^{θ}, θ_{ij}, and θ_{ij}^{0} are force constant, bond angle, and reference bond angle. K_{ij}^{r}, K_{ij}^{θ}, r_{ij}^{0} and θ_{ij}^{0} were derived from crystal structures, as well as from vibrational frequencies.[39,127,128] Recently, quantum chemistry computations have been increasingly used to derive these parameters.[41,45] Torsional energy is described by a sinusoidal term as shown in Equation 2.54. Atom i and j in the torsional term are separated by three bonds. V_{n}^{ij} is the force constant of a torsion, n is the multiplicity, ϕ_{ij} is the current torsional angle, and ϕ_{ij}^{0} is the phase offset (typically 0 or π). In CHARMM, n can be up to 6, while in AMBER and OPLS-AA, it is only considered up to 3 or 4.[57] It should be noted that a rotational barrier comes from both torsional energy, as well as from non-bonded (van der Waals and electrostatic) terms, therefore torsional parameters are highly correlated with the non-bonded parameters.[84] Torsional parameters are derived from experimental data or fitted to quantum chemistry data depending on different force fields.[34,127,129–133] The improper torsional terms are employed to avoid unphysical out-of-plane distortions of planar groups. Unlike the torsion angle-like term in Equation 2.55 used by OPLS-AA and AMBER, CHARMM employs a basic harmonic functional form for improper torsional energy as shown in Equation: 2.56:

$$E_{\text{improper}} = \sum_{improper}^{1-4atoms} K_{imp}^{ij} \left(\phi_{ij} - \phi_{ij}^{0}\right)^2, \tag{2.56}$$

where K_{imp}^{ij} is the force constant of an improper dihedral angle, ϕ_{ij} is the current angle, and ϕ_{ij}^{0} is the reference angle.

Nonbonded terms are typically limited to describing vdW interactions and electrostatic interactions. As shown in Equation 2.57 and 2.58, vdW interactions are described by a Lennard-Jones 6–12 term and electrostatic interactions are described by a Coulombic term.

$$E_{\text{vdW}} = \sum_{i<j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] f_{ij} \tag{2.57}$$

$$E_{\text{ele}} = \sum_{i<j} \frac{q_i q_j}{r_{ij}} f_{ij} \tag{2.58}$$

Nonbonded parameters include the LJ well-depth ε_{ij}, minimum energy distance σ_{ij}, and partial atomic charges q_i. Nonbonded interactions between third neighbors (1,4-interactions) are reduced

in most force fields with various scale factors f_{ij}. LJ parameters are assigned according to different atom types that represent atoms in a specific chemical environment. The LJ parameters between different atom types are calculated according to the combination rules.[134,135] For example, AMBER and CHARMM employ the Lorentz-Berthelot combination rule,[136] which uses the arithmetic mean for σ_{ij} and the geometric mean for ε_{ij}. While OPLS-AA employs the geometric mean for both σ_{ij} and ε_{ij}. LJ parameters can be obtained by fitting to experimental densities and heats of vaporization,[34] or deriving from quantum chemistry calculations with exchange-hole dipole moment (XDM) model[137] or atoms-in-molecule (AIM) method.[47] q_i can also be derived by fitting to experimental data, e.g. hydration free enthalpies in water,[40] or extracted from quantum chemistry data.[35]

Polarizable Force Field

Despite the many successes of classical FFs in MD simulations of biological systems, it is more and more clear that the inclusion of electronic polarization will play a central role in the next generation FFs.[81] In this section, we introduce several popular polarizable FF models.

Induced dipole model

The induced dipole model describes polarization energy by summing up the interactions between partial atomic charges and induced dipole moments.[69,138–140] Ponder *et al.* added the interactions between induced dipoles and higher permanent moments (up to quadrupoles) to the model.[74,141] In the induced dipole model, an additional energy term E_{pol} is added to the total energy. E_{pol} can be computed according to:

$$E_{\mathrm{pol}} = -\frac{1}{2}\sum_i \boldsymbol{\mu}_i \boldsymbol{E}_i^0, \tag{2.59}$$

where the summation is over all atomic sites, μ_i is the dipole induced on atom i, and E_i^0 is the electrostatic field at the polarizable site i generated by the current charge in the system. The factor of $1/2$ is a result of the induction cost for the formation of an induced dipole.[142] The induced dipole moment is proportional to the total electrostatic field, E_i:

$$\boldsymbol{\mu_i} = \alpha_i \boldsymbol{E}_i = \alpha_i \left(\boldsymbol{E}_i^0 - \sum_{i \neq j}^{N} \boldsymbol{T}_{ij} \boldsymbol{\mu}_j \right), \tag{2.60}$$

where the proportionality constant, α_i, is the polarizability of the atom i, T_{ij} is the dipole-dipole interaction tensor (dipole field tensor):

$$\boldsymbol{T_{ij}} = \frac{1}{r_{ij}^3}\left(1 - \frac{3\vec{\boldsymbol{r}_{ij}}\vec{\boldsymbol{r}_{ij}}}{r_{ij}^2}\right). \tag{2.61}$$

Most point dipole models only consider the interactions between static charges and induced dipoles to describe polarization effects. The AMOEBA model[74] considers interactions between charges, higher-order static atomic multipoles, and induced point dipoles. Therefore, Equation 2.60 is modified by treating the static electric field by permanent multipoles instead of by permanent atomic charges:

$$\boldsymbol{\mu_i} = \alpha_i \left(\sum_{j \neq i} \boldsymbol{T}_{ij}^{\alpha} \boldsymbol{M}_j + \sum_{k \neq i} \boldsymbol{T}_{ik}^{\alpha\beta} \boldsymbol{\mu}_k \right), \tag{2.62}$$

where M ($M_i = (q_i,\mu_{i,x},\mu_{i,y},\mu_{i,z},Q_{i,xx},Q_{i,xy},Q_{i,xz} \ldots Q_{i,zz})^T$) is the permanent atomic multipole component, and T_{ij} =[T_α, $T_{\alpha\alpha}$, $T_{\alpha\beta}$, $T_{\alpha\gamma}$, ... ,] α, β, $\gamma = x$, y, z is the interaction matrix.

Fluctuating Charge Model

The fluctuating Charge (FC) model introduces polarization effects by enabling the tuning of partial charges according to the electric field of their environment. It does not introduce any additional energy terms or particles to a classical FF. This can be performed based on electronegativity equalization: charges flow between atoms until electronegativity of the atoms become equalized.

Classical Drude oscillator model

The classical Drude oscillator model describes polarization effects by attaching a massless charged particle to each polarizable atom via a harmonic spring. For a given atom with charge q at the atomic center in the system, a Drude oscillator (or Drude particle) is introduced with a charge q_D assigned. The charge on the atom is adjusted to $q - q_D$ to represent the net charge of the atom-Drude oscillator pair. The Drude oscillator is attached to the atom harmonically with a force constant k_D. The position of the Drude oscillator is adjusted self-consistently to their local energy minima. The displacement, $\boldsymbol{d}$, of the Drude oscillator from the central atom can be calculated as:

$$\boldsymbol{d} = \frac{q_D \boldsymbol{E}}{k_D}, \tag{2.63}$$

where $\boldsymbol{E}$ is the local electric field.

The induced dipole, $\boldsymbol{\mu}$, is calculated as:

$$\boldsymbol{\mu} = \frac{q_D^2 \boldsymbol{E}}{k_D}, \tag{2.64}$$

which leads to a simple expression of polarizability, α,

$$\alpha = \frac{q_D^2}{k_D}. \tag{2.65}$$

Therefore, the only parameter in the classical Drude oscillator model is the combination of q_D^2 and k_D, which is responsible for the polarizability. It should be noted that the Drude model only uses the Coulombic term that already exists in a classical FF. No additional interaction terms are needed. Dipole-dipole interactions are balanced by additional charge-charge calculations. Therefore, there is no need to calculate the dipole field tensor T_{ij} in Equation 2.61, which greatly saves computational costs. **Paper II** in this thesis tested the Drude model using a quantum chemistry data set of cation-dipeptide systems.

CTPOL model

The CTPOL model combines charge transfer (CT) and polarization effects (POL) into classical FF. Instead of a fixed-charge model, CTPOL takes into account the charge transfer from the ligand atom L (O, S, N) to the metal cation. The amount of transferred charge, Δq_{L-Me}, is assumed to depend linearly on the interatomic distance, r_{Me-L}:

$$\Delta q_{(L-Me)} = a_L r_{Me-L} + b_L. \tag{2.66}$$

at distances greater than the sum of the vdW radii of atoms i and j, r_{ij}^{vdW}, charge transfer can be

neglected. Thus, the charge q_L on the ligand atom L can be calculated as:

$$q_L = q_L^0 + \Delta q_{(L-Me)}, \tag{2.67}$$

where q_L^0 refers to the charge on atom L from a given classical FF.

The polarization energy can be computed as shown in Equations 2.59 and 2.60. The summation is over the metal and the metal-bound atoms. A cutoff distance r^{cutoff}, which equals to the sum of the vdW radii of atoms i and j scaled by a parameter $\gamma = 0.92$, is introduced to avoid unphysically high induced dipoles at close distances. If the distance r^{ij} between atoms i and j is smaller than r^{cutoff}, we set r^{ij} equal to r^{cutoff}. The CTPOL model is tested in **Paper II** and implemented and parameterized in **Paper III** of this thesis.

2.2 Molecular Dynamics Simulation

Molecular dynamics simulations are a computer simulation technique in which the thermodynamic and kinetic properties of a set of interacting atoms can be calculated based on statistical mechanics. Nowadays, it has become one of the most important methods to study the microscopic interactions of biomolecules at the atomic level.[143,144] In MD simulations, the nuclei are approximated as classical particles and the dynamics are simulated by solving Newton's second equation:

$$\boldsymbol{F} = m\boldsymbol{a}. \tag{2.68}$$

The differential form of Equation 2.68 can be written as:

$$-\frac{dV}{d\boldsymbol{r}} = m\frac{d^2\boldsymbol{r}}{dt^2}, \tag{2.69}$$

where $\boldsymbol{r}$ contains the coordinates of all particles in the system, and V represents the potential energy at position $\boldsymbol{r}$. The equation is solved simultaneously in small time steps. The atomic positions and velocities along with time form the so-called trajectory. Atomic positions and velocities are information at the microscopic level. They describe the motion of particles and enable the calculation of macroscopic observables such as pressure, energy, and so on. The basis of such simulations is the integration of Equation 2.69. The predictive power of MD simulations is based on the ergodic hypothesis, which assumes that the average of a small number of particles over a long time is equivalent to the average of a large number of particles over a short time, i.e. time averaging is equivalent to ensemble averaging.[84]

2.2.1 Integrators

There are several algorithms for the numerical integration of Equation 2.69. Each algorithm has its advantages and disadvantages.

The Verlet algorithm

Commonly used algorithms are the Verlet algorithm and its variations.[145] The positions of particles, $\boldsymbol{r}_i$, after a small time step Δt can be calculated by Taylor expansion:

$$\begin{aligned} \boldsymbol{r}_{i+1} &= \boldsymbol{r}_i + \frac{\partial \boldsymbol{r}}{\partial t}(\Delta t) + \frac{1}{2}\frac{\partial^2 \boldsymbol{r}}{\partial t^2}(\Delta t)^2 + ... \\ &= \boldsymbol{r}_i + \boldsymbol{v}_i(\Delta t) + \frac{1}{2}\boldsymbol{a}_i(\Delta t)^2 + ..., \end{aligned} \tag{2.70}$$

where $\boldsymbol{v}_i$ is the velocity at time t_i, and $\boldsymbol{a}_i$ is the acceleration at time t_i. Similarly, the positions of particles a small time step Δt earlier can be written as:

$$\boldsymbol{r}_{i-1} = \boldsymbol{r}_i - \boldsymbol{v}_i(\Delta t) + \frac{1}{2}\boldsymbol{a}_i(\Delta t)^2 + ... \tag{2.71}$$

The trick with Δt and $-\Delta t$ allows to truncate the Taylor expansion. Combining Equation 2.70 with Equation 2.71 results in:

$$\begin{aligned} \boldsymbol{r}_{i+1} &= (2\boldsymbol{r}_i - \boldsymbol{r}_{i-1}) + \boldsymbol{a}_i(\Delta t)^2 \\ \boldsymbol{a}_i &= \frac{\boldsymbol{F}_i}{m_i} = -\frac{1}{m_i}\frac{dV}{d\boldsymbol{r}_i}. \end{aligned} \tag{2.72}$$

This is the classical *Verlet* algorithm[146] for solving Newton's equation. At the starting point, $\boldsymbol{r}_{i-1}$ is not available, but can be approximated from the Equation 2.70:

$$\boldsymbol{r}_{-1} = \boldsymbol{r}_0 - \boldsymbol{v}_0\Delta t. \tag{2.73}$$

The leap-frog algorithm

The velocity does not appear explicitly in the Verlet algorithm, which can be a problem for generating ensembles with constant temperature. This can be remedied by the *leap-frog* algorithm.[147]

Perform similar expansion of equations 2.70 and 2.71 with half a time step and then subtract:

$$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i + \boldsymbol{v}_{i+\frac{1}{2}}\Delta t. \tag{2.74}$$

The velocity is given as:

$$\boldsymbol{v}_{i+\frac{1}{2}} = \boldsymbol{v}_{i-\frac{1}{2}} + \boldsymbol{a}_i\Delta t. \tag{2.75}$$

Equation 2.74 and equation 2.75 are the leap-frog algorithm.

The velocity Verlet algorithm

In leap-frog algorithm, the velocity is explicitly present, which promotes the coupling to external heat bath. However, the position is always later than the velocity by half a time step. To eliminate this abnormality, the *velocity Verlet* algorithm[148] was introduced.

The equations of the velocity Verlet algorithm take the form:

$$\begin{aligned} \boldsymbol{r}_{i+1} &= \boldsymbol{r}_i + \boldsymbol{v}_i\Delta t + \frac{1}{2}\boldsymbol{a}_i\Delta t^2 \\ \boldsymbol{v}_{i+1} &= \boldsymbol{v}_i + \frac{1}{2}(\boldsymbol{a}_i + \boldsymbol{a}_{i+1})\Delta t. \end{aligned} \tag{2.76}$$

2.2.2 SHAKE and RATTLE

The above algorithms solve Newton's second equation by numerical integration. The time step is an important parameter for MD simulations. The smaller the time step Δt, the closer the trajectory is to the "true" trajectory. Typically, it is an order of magnitude smaller than the speediest mode in a system to get sufficient accuracy. The rotations and vibrations of a molecule usually occur with frequencies in 10^{11} to 10^{14} s^{-}, which means that the time step should be on the femtosecond (fs) level or less. However, many important reactions happen over a long time scale. For example, protein folding may happen in milliseconds or seconds. Simulating such a process is computationally expensive and requires to be performed by incredibly large number of time steps. Stretching vibrations, especially those containing hydrogen, are the fastest process for molecules. However, these motions have relatively little effect on molecular properties. Constraining bond lengths involving hydrogen atoms allows for larger time steps and therefore reduces the computational cost of the simulation. Typically, constraining bond lengths allows to increase the time step by a factor of 2 or 3.

Constraints can be done with SHAKE[149] or RATTLE[150] algorithms. The SHAKE algorithm is mainly applied in combination with the Verlet algorithm. After adding constraints to the Verlet algorithm, Equation 2.72 is expressed as:

$$\boldsymbol{r}_{i+1} = (2\boldsymbol{r}_i - \boldsymbol{r}_{i-1}) + \frac{(\boldsymbol{F}_i + g_s[\boldsymbol{r}_i, \boldsymbol{v}_i])}{m_i}(\Delta t)^2, \tag{2.77}$$

where g_s represents force due to the constraint. Velocity can be obtained from coordinates. The RATTLE algorithm was developed specifically for the Velocity Verlet algorithm. Similarly, under constraints, Equation 2.76 can be written as:

$$\begin{aligned} \boldsymbol{r}_{i+1} &= \boldsymbol{r}_i + \boldsymbol{v}_i\Delta t + \frac{(\boldsymbol{F}_i + g_{\mathrm{RR}}(t))}{2m_i}\Delta t^2 \\ \boldsymbol{v}_{i+1} &= \boldsymbol{v}_i + \frac{1}{2m_i}(\boldsymbol{F}_i + \boldsymbol{F}_{i+1} + g_{\mathrm{RR}}(t) + g_{\mathrm{VR}}(t))\Delta t, \end{aligned} \tag{2.78}$$

where g_{RR} and g_{VR} are two separate approximations related to the constraint force.

2.2.3 Molecular Dynamics Ensembles

Coordinates and velocities are microscopic properties of a system. They can be obtained by the integration algorithms described above. The macroscopic properties of a system include volume (V), pressure (P), and temperature (T). The concept of *ensemble* connects the microscopic properties of a system with the macroscopic properties. An ensemble is a collection of many microstates but have an identical macrostate.[151]

Simple MD simulations produce micro-canonical ensemble (NVE). NVE ensemble corresponds to an isolated system where the number of particles (N), volume (V), and energy (E) are fixed, while temperature (T) and pressure (P) are fluctuating. The total energy of a system has two

parts: kinetic energy and potential energy. It can be expressed as:

$$\begin{aligned} E_{\text{tot}} &= E_{\text{kin}} + E_{\text{pot}} \\ &= \sum_{i=1}^{N} \frac{1}{2} m_i \boldsymbol{v}_i^2 + E_{\text{pot}}. \end{aligned} \tag{2.79}$$

The temperature and pressure can be calculated from:

$$\begin{aligned} E_{\text{kin}} &= \frac{3}{2} N k_B T \\ &= \frac{3}{2} PV \end{aligned} \tag{2.80}$$

where k_B is the Boltzmann constant.

Typically, a NVT or NPT ensemble represents the reality of experiments better than a NVE ensemble. Since the temperature is proportional to the average kinetic energy, it can be controlled by scaling the velocities at each time step. One example is the Berendsen thermostat.[152] The Berendsen thermostat method couples the system to a "heat bath" that transfers energy to the system. The rate of energy transfer is controlled by a parameter τ:

$$\frac{dT}{dt} = \frac{1}{\tau}(T_{\text{desired}} - T_{\text{actual}}). \tag{2.81}$$

The velocity scale factor, s_v, is given by:

$$s_v = \sqrt{1 + \frac{\Delta t}{\tau} \left(\frac{T_{\text{desired}}}{T_{\text{actual}}} - 1 \right)}. \tag{2.82}$$

Another widely used method is the Nosé–Hoover thermostat.[153] In the Nosé–Hoover thermostat method, the heat bath is taken as a part of the system and fictive dynamic variables are assigned. The fictive dynamic variables change on the same footing with other variables.

The pressure can be maintained similarly by coupling to a "pressure bath". In Berendsen barostat method,[152] the coordinates are scaled to change the volume of the system instead of scaling velocities:

$$\begin{aligned} \frac{dP}{dt} &= \frac{1}{\tau}(P_{\text{desired}} - P_{\text{actual}}) \\ s_c &= \sqrt[3]{1 + \kappa \frac{\Delta t}{\tau} (P_{\text{actual}} - P_{\text{desired}})}, \end{aligned} \tag{2.83}$$

where s_c is the scale factor of coordinates, and κ represents the compressibility of the system. The pressure can also be held constant by the Nosé–Hoover method.

2.3 Optimization and Search

Many computational chemistry problems can be categorized as optimization problems in multi-dimensional space. Optimization means finding a minimum point in a search space. But in many cases, there are numerous different minima in a multi-dimensional space. The minimum with the lowest value is the *global* minimum, while the others are all *local* minima. The interest may lie in the global minimum or local minima in different cases. Optimization is a gigantic field. In this thesis, we employ several optimization methods for different optimization problems:

- The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [154] in Section 2.3.1 is used to perform geometry optimization on the Born-Oppenheimer potential energy surface. In this case, the local minima are of interest.

- Flexible biomolecules have a large structure space. Proper conformational sampling methods are often required to find the global minimum of the system. Two global optimization and search methods, genetic algorithm (GA) in Section 2.3.2, and replica exchange molecular dynamics (REMD) in Section 2.3.4, are employed to facilitate the conformational sampling.

- To determine the set of parameters in a force field, if the parameters enter the function in a quadratic way, e.g. ε_{ij}, the optimization of the function can be done by solving a set of linear equations. In this case, regularized linear regression in Section 2.3.5 is employed.

- If the force field parameters don't enter the function in a quadratic way, e.g. charge transfer parameters a_L and b_L, the global optimization method particle swarm optimization (PSO) in Section 2.3.3 is used.

2.3.1 Geometry optimization

There are several standard local optimization tools to deal with the task of geometry optimization as described in reference. [155] The molecular simulation code FHI-aims [123] employs a slightly different approach, which is reviewed in the following.

To find the local minima, the first and second derivatives with respect to atomic positions, which are the force and Hessian matrix, are required. Assume x_n is the set of atomic positions of a system at the optimization step n. The corresponding force and Hessian matrix of the system are:

$$\begin{aligned} f_n &= -\frac{\partial E}{\partial x_n} \\ H_n &= \frac{\partial^2 E}{\partial x_n^2}. \end{aligned} \tag{2.84}$$

To determine that a point on the PES is a local minimum requires that $f_n = 0$ and H_n be positive semidefinite. If the values of $f_n = 0$ and H_n are known, the PES around the local minima can be written as second order Taylor expansion in a displacement s_n as:

$$M(x_n + s_n) = E_n - f_n^T s_n + \frac{1}{2} s_n^T H_n s_n. \tag{2.85}$$

The calculation of the exact Hessian is computationally expensive. However, it is not necessary to be known. One can use an approximate matrix to replace the exact Hessian matrix and update it during the optimization. The most widely used approach to update the estimated matrix is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm: [154]

$$H_{n+1} = H_n - \frac{H_n \Delta x_n \Delta x_n^T H_n}{\Delta x_n^T H_n \Delta x_n} - \frac{\Delta f_n \Delta f_n^T}{\Delta f_n^T \Delta x_n}, \tag{2.86}$$

where $\Delta x_n = x_{n+1} - x_n$ and $\Delta f_n = f_{n+1} - f_n$. The initial guess of the matrix is important in this method and dramatically affects the efficiency of the optimization process. A different initial

matrix may even affect the final results in some cases.[156] The naive choice for the initial matrix is to take the scaled identity matrix:

$$H_0 = \beta I, \tag{2.87}$$

where $\beta > 0$.

2.3.2 Genetic Algorithm

Genetic Algorithm (GA) is employed in this thesis to perform conformational sampling on the high-dimensional PES. GA is one of the most popular proposed evolutionary algorithms, and it is frequently used for global structural sampling of molecules.[157–159] GA operators are mainly selection, crossover, and mutation. GA follows the concept of survival of the fittest. Each solution is like a chromosome, and each parameter corresponds to a gene. The fitness of solutions in a population is evaluated by a fitness (objective) function. The GA algorithm starts with a pool of random populations. Good solutions are selected while poor solutions are removed from the pool. Selected solutions are employed to create new generations. Two solutions (parent solutions) are combined to produce two new solutions (offspring solutions) by crossover. Different crossover techniques can be employed.[160–162] One or multiple "genes" are altered in the newly created offspring solutions to introduce another level of randomness.

GA-based structural sampling combined with local optimization is easy to implement, stable, no need to assume the importance of specific degrees of freedom, and does not need to provide structural preferences. It has been proven to be robust for locating points close to the global minimum. The GA-based structural sampling in this thesis is performed by the flexible, open-source package Fafoom.[163] Firstly, genetic algorithm at the PBE+vdw^{TS} level with *light* basis is employed to perform the structural sampling. Then a clustering scheme with clustering criterion of 0.02 for RMSD of atomic positions and 0.02 kcal/mol for relative energy is applied to remove duplicates. The obtained conformers are further relaxed using FHI-aims at the PBE+vdW^{TS} level with tight basis sets. Final conformers are obtained after clustering.

2.3.3 Particle Swarm Optimization

Particle swarm optimization (PSO)[164] is a population-based algorithm and belongs to the group of swarm-based algorithms. It mimics the navigation mechanism of birds in nature. In PSO, each solution is considered as a "particle" that can move on a search landscape. Each particle adjusts its position based on its historical positions and those of other particles. During the movement, position vector and velocity vector are needed. The position vector ($\boldsymbol{X}$) represents the value of the interested problem, in our case, it's a set of force field parameters, and the velocity vector ($\boldsymbol{V}$) represents the direction and speed of the movement.

The position vector of each step is defined as:

$$\boldsymbol{X}_i(t+1) = \boldsymbol{X}_i(t) + \boldsymbol{V}_i(t+1), \tag{2.88}$$

where $\boldsymbol{X}_i(t)$ is the position of ith particle at tth iteration, and $\boldsymbol{V}_i(t+1)$ is the velocity of ith particle at the iteration t.

The velocity vector is defined as:

$$\boldsymbol{V}_i(t+1) = w\boldsymbol{V}_i(t) + c_1 r_1 \left(\boldsymbol{P}_i(t) - \boldsymbol{X}_i(t)\right) + c_2 r_2 \left(\boldsymbol{G}(t) - \boldsymbol{X}_i(t)\right), \tag{2.89}$$

where $\boldsymbol{V}_i(t)$ is the velocity of ith particle at tth iteration, w is the inertia weight, c_1 and c_2 are the individual coefficient and social coefficient, respectively, r_1 and r_2 are random numbers between $[0, 1]$, $\boldsymbol{P}_i(t)$ and $\boldsymbol{G}(t)$ are the best solutions found by the ith particle and all particles until iteration t, respectively, and $\boldsymbol{X}_i(t)$ is the position of ith particle at iteration t.

In Equation 2.89, the first term considers how much of the previous velocity should be maintained. The second term is the so-called cognitive component, which represents the individual intelligence. $\boldsymbol{P}_i(t)$ is the best position of the particle so far. The third term represents the social intelligence. Therefore, the movement of each particle is affected by individual and social intelligence and by that dragged towards the best regions of search space. The PSO starts with a random set of particles. Each particle has a random position vector and velocity vector. Parameters of w, c_1 and c_2 are initialised. Then, particles move in different directions in the search space. Positions of particles are updated until the end condition is satisfied.[165]

2.3.4 Replica-exchange Molecular Dynamics

The original replica-exchange method[166] was applied with Monte Carlo simulations.[167–169] It was then combined with molecular dynamics and formed the method of replica-exchange molecular dynamics (REMD).[170] The basic idea of REMD is to run MD simulations of multiple replicas of the system under study simultaneously. Each simulation is in the canonical ensemble at different temperatures. Thus, the replica-exchange method is also called parallel tempering. Figure 2 shows the schematic representation of REMD method. REMD is not an optimization technique per se, but in this thesis, we use it to explore the structural space of bio-molecules. The combination of REMD, clustering, and local optimization forms a global search technique.

Standard MD simulations are of limited use when performing conformational sampling on potential energy surfaces with many minima. Due to the inflexible nature of MD, sampling can easily be trapped in metastable minima on the surface, resulting in incomplete sampling. At high temperatures, barriers are easier to overcome and the trajectory is less likely to be trapped in local minima. While the local energy surface around local minima can be accurately sampled at low temperatures. The REMD method combines these two advantages by swapping replicas at different temperatures after a specific MD simulation time. In this way, a random walk through a predefined discrete temperature space is introduced, which enables the simulation to overcome the energy barrier and sample a wider space. In REMD simulations, swaps occur between replicas with neighboring temperatures. When two replicas are swapped, not only atomic positions but also their momenta are swapped. To adapt the momenta of the replicas to new temperature, the easiest

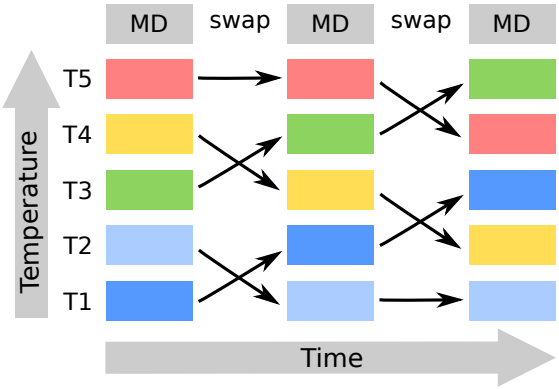


Figure 2: The schematic representation of replica-exchange molecular dynamics method. Five replicas swap between five different temperatures. MD simulations are performed in-between the swaps.

way is to scale the momenta according to the approach proposed by Sugita and Okamoto:[171]

$$p_i' = \sqrt{\frac{T_{\text{new}}}{T_{\text{old}}}} p_i, \tag{2.90}$$

where p_i are the old momenta of particle i, T_{old} and T_{new} are the temperatures before and after the swap, respectively. In this way, the average kinetic energy is preserved as $\frac{3}{2}Nk_BT$. The probability of accepting a swap between ensembles i and j is determined by the Metropolis acceptance criterion:[172]

$$P_{\beta_i \leftarrow \beta_j} = min\left[1, e^{-(\beta_j - \beta_i)(E(\mathbf{R}_i) - E(\mathbf{R}_j))}\right], \tag{2.91}$$

where $\beta_i = 1/k_BT_i$, and $E(\mathbf{R}_i)$ is the potential energy of the ensemble i with configuration R.

2.3.5 Regularized Linear Regression: Ridge Regression and LASSO

Multiple linear regression is to investigate the relationship between two or even more independent variables ($\boldsymbol{x}_i$) and one dependent variable ($\boldsymbol{y}_i$). For a multiple linear regression model, assume a data set contains n samples, $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2)$, $\ldots$, $(\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_i$ is the input vector containing p predictor variables, *i.e.*

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}, \tag{2.92}$$

and y_i is the i-th output value. The multiple linear model specifies a linear relationship between $\boldsymbol{x}_i$ and the expected value $\hat{y}_i$, that is,

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, i = 1, \ldots, n, \tag{2.93}$$

where β_d, $d = 0, 1, \ldots, p$ represent linear regression coefficients. If we define the coefficient vector $\boldsymbol{\beta}$ as:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \tag{2.94}$$

vector $\boldsymbol{Y}$ as:

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \tag{2.95}$$

and matrix $\boldsymbol{X}$ as:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \tag{2.96}$$

we can obtain Equation 2.97 from Equation 2.93

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta}. \tag{2.97}$$

The regression coefficients can be obtained by the method of least squares.[173] With the property of matrix, Equation 2.97 can be written as:

$$\boldsymbol{X}^T\hat{\boldsymbol{Y}} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}. \tag{2.98}$$

Thus, assuming the rank of the matrix $\boldsymbol{X}$ equals to $p + 1$ so that $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is well defined, the least squares estimators $\hat{\boldsymbol{\beta}}$ can be obtained by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.99}$$

Equation 2.97 can be further written as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.100}$$

Thus the residual sum of squares (RSS) can be obtained by

$$\begin{aligned} RSS &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ &= \boldsymbol{Y}^T \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\boldsymbol{Y}, \end{aligned} \tag{2.101}$$

where $\boldsymbol{I}$ represents the identity matrix.

The matrix $\boldsymbol{X}^T\boldsymbol{X}$ may be ill-conditioned, or even singular if there are high correlations among predictor variables or a large number of predictors ($p > n$), which means that the inverse matrix of $\boldsymbol{X}^T\boldsymbol{X}$ is numerically unstable, thus further leading to numerically unstable least squares estimates $\hat{\boldsymbol{\beta}}$.[174] Various regularization or shrinkage regression[175] techniques can be used to address this

problem. These techniques impose a constraint on the model parameters, which 'shrinks' the regression coefficients towards zero, aiming to stabilize them. This can be achieved by adding a penalty term, $\mathscr{F}(\beta_0, \ldots, \beta_p)$, to the objective function that needs to be minimized

$$\mathscr{D} = \sum_{i=1}^{n} \left[Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right]^2 + \mathscr{F}(\beta_0, \ldots, \beta_p). \tag{2.102}$$

Ridge regression[176,177] employs a quadratic form for the penalty function, *i.e.* square of the magnitude of the coefficients. Thus, the objective function can be written as

$$\mathscr{D} = \sum_{i=1}^{n} \left[Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right]^2 + \lambda \sum_{l=1}^{p} \beta_l^2, \tag{2.103}$$

where λ represents a regularization or tuning parameter. λ acts as the Lagrange multiplier of the constraint. Equation 2.103 is equivalent to

$$\mathscr{D} = \sum_{i=1}^{n} \left[Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right]^2 \text{ subject to } \sum_{l=1}^{p} \beta_l^2 \leq \lambda, \tag{2.104}$$

and can be written in matrix notation as following

$$\mathscr{D} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}. \tag{2.105}$$

Minimizing $\mathscr{D}$ gives the equation[178]

$$(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{Y}, \tag{2.106}$$

where $\boldsymbol{I}$ denotes the identity matrix. Thus, the ridge regression estimates $\hat{\boldsymbol{\beta}}$ are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \tag{2.107}$$

Compared to Equation 2.99, a diagonal "ridge" ($\lambda \boldsymbol{I}$) is added to $\boldsymbol{X}^T \boldsymbol{X}$ matrix so that it always has an inverse, which further stabilizes the ridge regression estimates $\hat{\boldsymbol{\beta}}$.

We can see that when $\lambda \to 0$, the objective function $\mathscr{D}$ becomes similar to the linear regression objective function, while the larger its value, the stronger the coefficients' size penalized from Equation 2.107. Hoerl and Kennard[179] showed that there are always some λ (> 0) such that the residual sum of squares RSS is smaller than for the ordinary least squares estimators.

The penalty term in Ridge regression is the so-called "L_2" penalty, one can also use the "L_1" penalty in the objective function $\mathscr{D}$, which results in the LASSO (least absolute shrinkage and selection operator)[180] model:

$$\mathscr{D} = \sum_{i=1}^{n} \left[Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right]^2 + \lambda \sum_{l=1}^{p} |\beta_l|. \tag{2.108}$$

Unlike Ridge regression, there is no clear expression for the estimator $\hat{\boldsymbol{\beta}}$, but a number of solutions are required. The use of "L_1" penalty sets some of the estimators to 0, meaning that only a subset of the original estimators is obtained, which is the so-called sparse solution. LASSO regularization does not only act as shrinkage regression, but also a variable selector, it suppresses the low-impact predictor variables to 0. Although Ridge regression usually has higher predictive capability when

there is high multicollinearity among predictor variables, LASSO regression is often used for large number of predictor variables or overdetermination.[178]

The tuning parameter λ in Ridge regression and LASSO regression is often chosen by k-fold cross-validation approach.[181] One advantage of cross-validation is that it reduces over fitting without saving a subset of the data set for internal validation.[182]

2.4 Data Management

The data generated in this thesis has great potential for force field parameterization and further applications, e.g. machine learning or benchmarking. To make the data available to experts in force field development, or even to experts in other scientific fields, the data storage should fulfill the so-called FAIR principles:[183] findable, accessible, interoperable, and re-usable.

- "Findable" means that the data and corresponding meta-data should be easy to find. Metadata is the so-called data about data. It is a set of attributes necessary to annotate, characterize, and ultimately reproduce data. For example, a DFT total energy calculated with a specific functional and atomic positions. The total energy is seen as data, while the functional and atomic positions are the input to the calculation, necessary to reproduce the data, and therefore metadata for that calculation. The total energy can also be meta-data for further analysis.
- "Accessible" requires open and free authentication and authorization protocols.
- "Interoperable" requires the data to be able to integrate with other data, different applications or workflows.
- "Re-usable" is the biggest benefit of FAIR data handling. The data should be easy to apply to other applications or workflows.

Several repositories have been developed to store and share data. Some of them use metadata schemes to annotate data and make the data accessible through application programming interface (API). Furthermore, Semantic Web technologies are developed to enable a machine to "understand" the data by enriching the web with machine-processable information.[184] Ontologies are most closely related to the development of the Semantic Web. Several repositories, the concepts of Semantic Web and ontologies are outlined in the following.

2.4.1 Repositories

There are several repositories available for storing molecular computational data. To name some of them, ioChem-BD platform[185] supports data curation, storage, indexing data, and search engine services. It supports 11 computer codes. Complete input and output data are not stored in ioChem-BD. Some results in ioChem-BD are translated into CML (Chemical Markup Language), a chemistry-oriented XML language, and can be downloaded. The Benchmark Energy & Geometry

Database (BEGDB)[186] contains highly accurate QM calculations of molecular structures, energies, and properties. The data is listed in a table. Complete input and output data are also not available. The Novel Materials Discovery (NOMAD) Repository & Archive is explained in more detail as it is used to store data in this thesis.

NOMAD Repository & Archive supports over 40 computer codes and stores over 11 million entries.[187] It contains complete input and output files, so that the calculations can easily be repeated or continued. Each entry/calculation has a unique identifier. In addition, undivided entries can be curated into data sets for which a DOI can be assigned to make the data citable. In addition to raw data, a meta-data schema NOMAD Metainfo[188] is used to annotate and structure data. Data as well as meta-data from electronic-structure theory, e.g., structure, energy, program, are well organized in NOMAD Metainfo in a standardized, code-independent format. Information in NOMAD Metainfo can be explored through an API. The widely adopted JSON format is used to store data, which improves the interoperability of the data.

The NOMAD Repository & Archive fulfills the FAIR principles. The meta-data schema and unique paths for each term make the data finable (F). The API connected to the meta-data schema ensures accessibility (A). Using widely processable format like JSON to store data improves interoperability (I). Storing all raw data makes data better re-usable (R). The ontologies developed in this thesis aim to further improve the accessibility and interoperability.

2.4.2 Semantic Web

Before introducing the concept of ontologies, we need to describe the concept of the Semantic Web. As the definition from Berners-Lee et al.,[189] Semantic Web *"is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*. It refers to the vision of an intelligent network. Tim Berners-Lee presented his plan for Semantic Web Architecture at the XML 2000 conference[190] (Figure 3). Items in Figure 3 are

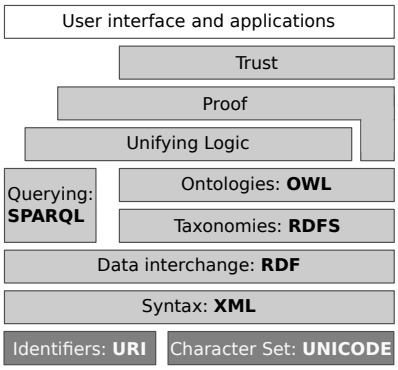


Figure 3: Semantic Web layer cake, based on Berners-Lee[190] and Reference.[191]

briefly introduced in the following:

- **Uniform Resource Identifier (URI)** and its extension, the internationalized resource identifier (IRI), represent unique entities.

- **Unicode** refers to the text standard expressed in most of the world's writing systems. For example, the latex source of this thesis is written in unicode UTF-8.
- **Extensible Markup Language (XML)** is a standard syntax to serialize and store data. It enables users to build human-readable as well as machine readable documents in a tree structure.[192] XML tags contain meta-data and represent the data structure, while it does not impose semantic restrictions on the meta-data in these documents.[193]
- **The Resource Description Framework (RDF)** is used for encoding, exchange, and reuse of metadata. The information about resources is represented in graph form. The RDF description is based on the triple relation: *subject-predicate-object*. Subject and predicate are usually resources identified uniquely by URIs or IRIs, while object can be a literal. Unlike XML, no assumptions about data structures are required in RDF.
- **RDF Schema (RDFS)** is an extension of RDF. It provides classes and properties and therefore provides the basic building blocks of concepts.
- **Web Ontology Language (OWL)** extends the RDFS and is syntactically embedded in RDF. OWL provides additional vocabulary to express relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), characteristics of properties (e.g. symmetry), and much more. Thus it provides greater machine interpretability than XML, RDF, and RDFS.[193]
- **SPARQL** is a query language. It can be used to query RDF, RDFS, and OWL.

To make sure that all results are trustworthy for users or applications, all semantics should be proofed and only based on trusted inputs.

2.4.3 Ontologies

Ontology is a word borrowed from philosophy. In philosophy, ontology is the study of being - about what kind of things exist. Nowadays, it is borrowed by computer science as a semantic knowledge organization system (KOS). A widely cited description of ontologies is made by Gruber,[194] which is that *"An ontology is an explicit specification of a conceptualization."* Later, Studer et al.[195] updated the definition as *"An ontology is a formal, explicit specification of a shared conceptualization, where 'formal' means the ontology should be machine readable; 'explicit' requires all concepts, properties, relations, functions, constraints, and axioms to be explicitly defined; 'shared' emphasizes that the ontology represents consensual knowledge, e.g. that it is accepted by a group; and 'conceptualization' is an abstract model of some phenomenon in the world."* However, the definition is still rather abstract. To make it easier for beginners to understand: an ontology defines a common vocabulary for researchers who need to share information by including machine-interpretable definitions of basic concepts in a domain and relations among them.[196]

The main components of an ontology include classes, properties, instances, and axioms. Classes are the focus of most ontologies. They represent concepts of entities in a domain. A class can have

sub-classes that describe more specific concepts and a set of individual instances. Properties allow users to assert general facts about classes and specific facts about individuals. They also capture relationships between classes. Like classes, properties can have sub-properties. Once the classes and properties are defined, an ontology is produced. Instances are the "thing" that a class represents. For example, "zinc" is an instance of class "atom". Strictly speaking, as a conceptualisation of a domain, an ontology should not contain any instances. To further understand this, two concepts need to be distinguished: ontology and knowledge graph. There are two types of statements in computer science. The *terminological component* consisting of TBox statements defines classes and properties. The TBox statements build a conceptual framework to express actual facts. The facts or data are represented with the vocabulary defined by TBox in so-called *assertion components*. The *assertion component* consists of ABox statements, and instances belong to the ABox. In summary, the TBox is usually an ontology and the ABox contains the facts or data stored in the schematic definitions by the TBox. TBox and ABox together form a knowledge graph composed of the ontology and the data. However, sometimes it's difficult to define things as classes or instances. For example, one can state "zinc" is an instance of class "atom". It could be argued that "zinc" is a class that represents different instances of zinc, e.g. Zn^0 and Zn^{2+}. This is a well-known open question of ontology management. Finally, axioms are the facts in an ontology.

An ontology is developed to share a common understanding of knowledge in a domain among people or software agents. Sharing data with ontology also fulfills the FAIR principles. Ontologies ensure that data and meta-data are both human-readable and machine-readable, thereby enabling automatic discovery of data (F). In principle, any question framed in ontology logic can be answered in finite steps by ontology query language. This ensures accessibility (A). Ontologies ensure interoperability (I) by presenting data in a formal language and format. Ontologies can be easily connected to other ontologies or applied to other applications, making the data re-usable (R). Through these principles, data is extensible, accessible, and automatically processed.

There is no correct way to develop an ontology. Typically, developing an ontology is an iterative process. Developers start with a rough ontology, then revise and refine the ontology iteratively by using it in applications or discussing with experts. Ontologies can be developed in two ways: bottom-up or top-down. The bottom-up approach starts with an existing database and extends the specific concepts in the database to connect upper-level concepts. This is a practical way to present and explore data quickly. After the data is populated, a knowledge graph is built. The top-down approach starts with the most general concepts and properties, regardless of the existing database. Both approaches have their advantages and users can choose which approach to use based on their purpose. The two approaches can also be used in combination.

There are several typical steps in developing ontologies. Before developing an ontology, the scope and usage of the ontology need to be determined. The scope of an ontology can be defined by considering which questions the ontology should be able to answer. Then, one should check for existing ontologies that can be reused. Reusing existing ontologies may be necessary if one wants to link their ontologies with other ontologies. The next step is to write down important terms

and define classes and the hierarchy. Once the classes and hierarchy are defined, one can define properties and restrictions on properties. Classes are connected to each other through properties. Finally, individual instances are created.

Chapter 3

Summary of main results

This chapter summarizes the main results of the publications that make up this thesis. Details of methodologies and secondary results are not included in this chapter and can be found in the respective publications in Chapter 4. The main purpose of the work underlying this thesis is to pave the way for accurate simulation of metalloproteins. To that end, we performed a systematic study and presented the results through three papers.

- **Paper I**[82] presents a quantum chemistry data set of amino-methylated and acetylated (capped) dipeptides with possible protonation states and several divalent cations (Ca^{2+}, Mg^{2+} and Ba^{2+}) that can be used for FF parameterization.
- **Paper II**[83] benchmarks a polarizable Drude FF, and three widely used classical FFs, OPLS-AA, AMBER, and CHARMM (C36) against the quantum chemistry data set from **Paper I**, and demonstrates how QM-driven parameterization and the explicit consideration of charge transfer and polarization effects can improve the simulation of cation-protein interactions.
- **Paper III** presents an open source parameterization tool, which enables the parameterization of classical FF OPLS-AA as well as CTPOL model, which is a FF model that includes charge transfer and polarization effects.

Classical force field parameterization has been used to improve the accuracy of metalloprotein simulations, and has been successful to some extent.[50,56,72] However, it is still limited due to the complex electrostatic environment in metalloproteins. An alternative approach is to introduce more physics, e.g. charge transfer and polarization effects, to the FF framework. Studies have shown that polarizable FFs better reproduce experimental as well as high-level quantum chemistry results than classical FFs.[65,197] Explicit inclusion of charge transfer and polarization effects plays an important role in the development of next-generation force fields. However, more complex functions or more energy terms result in more parameters, which makes the parameterization even more challenging. Thus, large and sufficiently accurate data sets are needed. The reference data for force field parameterization can be experimental data or computational data. However, experimental data of ion-containing systems are limited and lack less stable conformations. Due

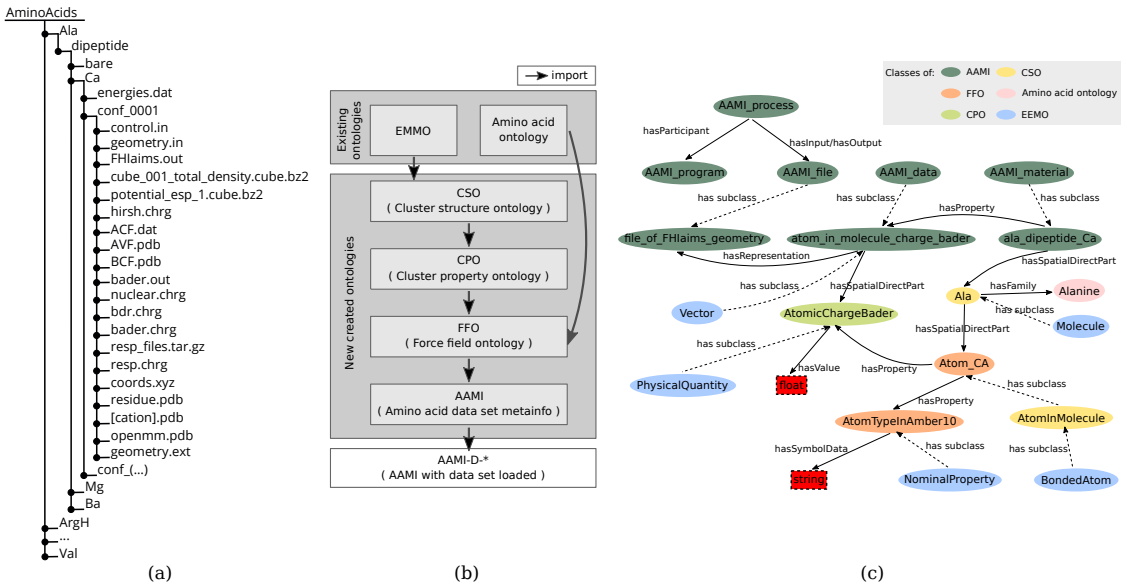


Figure 4: (a) Hierarchy of the ontologies linked to amino acid data set meta-info (AAMI). (b) Partial high level classes of AAMI ontology. Ovals and Rectangles represent classes and literals, respectively. Classes in different ontologies are labeled with different colors. Solid lines represent properties, and dashed lines indicate the property of "*has subclass*".

to the good accuracy and affordable computational cost, DFT data has been used for force field development[198] and has been shown to improve force field accuracy.[199] Although several studies have provided a solid basis for conformational and energetic assessment of protein building blocks, these data vary in approximations and sampling methods. Furthermore, data is often not available in a straight forward usable way.

We believe that better FF start with better data. Hence we provide a DFT data set that covers a wide range of amino-methylated and acetylated (capped) dipeptides to simulate the diverse chemical spaces in proteins in **Paper I**.[82] The data set contains:

- 20 proteinogenic amino acids including possible protonation states of side chains;
- three divalent cations (Ca^{2+}, Mg^{2+}, and Ba^{2+});
- 21,909 stationary points on the PES with a wide range of relative energies up to 4 eV;
- properties related to force fields development, such as partial charges related to electrostatic interactions, interaction energies related to cation-protein interactions, and so on.

All data are calculated on consistent computational footing. The reliability of the xc-functional and sampling method employed was evaluated in previous studies.[105,163] These characteristics make the data set suitable for various uses, such as force field development, machine learning, and force field benchmarking. To make the data available to the community, we shared the data set via the NOMAD Repository & Archive. However, only electronic structure theory data and the corresponding meta-data are represented in the NOMAD archive. The conceptual knowledge about force field related secondary data of our study such as atom type and connectivity is hidden

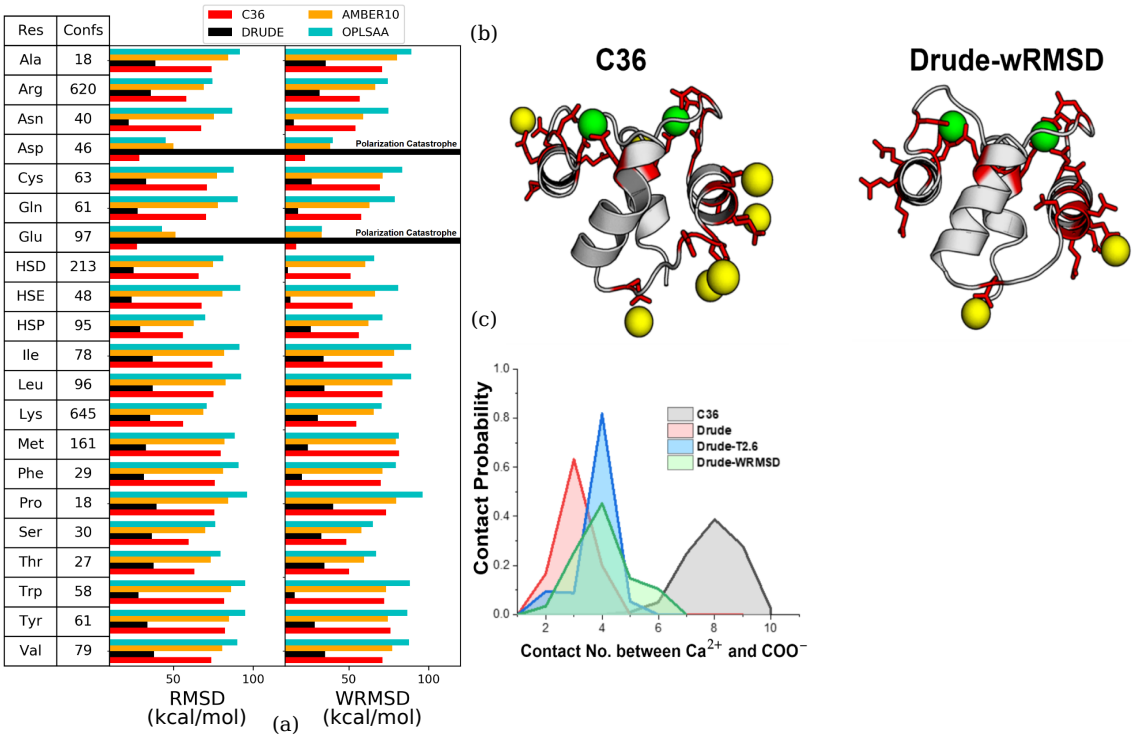


Figure 5: (a) Number of conformers, RMSD as well as Boltzmann-weighted RMSD (wRMSD) of the four FFs relative to the quantum chemistry interaction energies for each dipeptide-Ca^{2+} system. RMSD and wRMSD are in kcal/mol. (b) Characteristic snapshots of N-lobe of the human calmodulin (CaM) protein in simulation with CHARMM36 (left) and Drude-wRMSD (right) models, respectively. Green spheres in the figure represent the Ca^{2+} ions binding with sites in Loop I and II. Gold spheres represent the Ca^{2+} ions recruited from solutions. (c) The distribution of the contacts between Ca^{2+} and COO^{-} with different FF models.

somewhere in the files, which hinders the automatic access and processing of the data. To alleviate this, several ontologies were developed to represent the data set. An ontology defines a machine-readable common vocabulary of concepts in a domain and relations between them. Figure 4 (a) shows the hierarchy of the developed ontologies. Two existing ontologies, European Materials Modelling Ontology (EMMO) and Amino Acid Ontology, were reused. In this way, Amino Acid Meta-Info (AAMI) is connected to upper-level concepts and is easily linked to more ontologies. Some of the high level classes and properties of AAMI are shown in Figure 4 (b) to give an overview of how AAMI is organized and how concepts are linked to each other. In principle, all questions in the ontology logical framework can be answered in several steps. Data queries can be done using the query language SPARQL. This query and answer framework further makes the data interoperable and re-usable.

To assess the reliability of force field models in describing cation-dipeptide interactions, the DFT data set in **Paper I** was used to benchmark four force field models in **Paper II**. The four force field models include one polarizable Drude model and three classical models, CHARMM (C36),[124] AMBER (AMBER10),[129] and OPLS-AA.[126] Figure 5 (a) displays the RMSDs as well as

Boltzmann-weighted RMSDs (wRMSDs) between the four FFs and quantum chemistry interaction energies for different dipeptide-Ca^{2+} systems. Boltzmann-weighted RMSDs put more weight on low-energy conformations during the evaluation. Clearly, the Drude model is more accurate than the other three non-polarizable FFs for almost all dipeptide-Ca^{2+} systems. For two systems, Glu-Ca^{2+} and Asp-Ca^{2+}, we observe unphysically large energies when two polarizable atoms come too close to each other. This so-called polarization catastrophe occurs between Ca^{2+} and oxygen atoms in our test set and can be remedied by implementing the Thole damping factor.

With the optimized pair-wise Thole parameter and LJ parameters, the Drude models no longer exhibit the polarization catastrophe phenomenon and yield much smaller RMSDs of interaction energies than the three non-polarizable FF models. It was also found that in addition to the Thole parameter, the optimization of the LJ parameters is also essential to better reproduce the DFT interaction energies. The optimized Drude parameters were then evaluated by MD simulations with the N-lobe of the human calmodulin (CaM) protein. Studies have shown that non-polarizable FFs overestimate the coordination number (CN) between Ca^{2+} and CaM protein.[52] The number of Ca^{2+} cations binding to a single CaM protein can be up to 20 with non-polarizable FF instead of the expected 4.[52] Figure 5 (b) shows the snapshots from the MD simulations of CaM protein with CHARMM36 and optimized Drude model (Drude-wRMSD). The probability distributions of the Ca^{2+} – carboxylate CNs are shown in Figure 5 (c). The average CN of the CHARMM36 simulation is 8.5, while all Drude simulations yield a smaller average CN of 4. This work showed how a combination of a large and comprehensive quantum chemistry database and condensed-phase MD simulations can drive force field development to simulate important metalloproteins.

Although the Drude model has shown its potential to better simulate metalloproteins, it may still have its limitations when charge transfer effects are significant. Studies have shown that the charge perturbation on the ligand atoms caused by Ca^{2+} is significant, and this impact occurs not only in the first coordination shell, but also in the second shell.[200,201] The CTPOL model incorporates charge transfer and polarization effects into the classical FF formula. **Paper II** tested the CTPOL model with Glu-Ca^{2+} and Asp-Ca^{2+} systems. The results showed that the inclusion of polarization effects can better reproduce the DFT interaction energies than classical FFs, while the introduction of charge transfer effects can further improve the accuracy of FF. It was also found that tuning the LJ parameters is critical after extending the standard FF to the CTPOL model.

Based on the fact that most polarizable FFs are subject to limited validation, they may need to be re-parameterized when they are used on new systems. However, re-parameterization is always time-consuming and tedious. Especially for the parameterization of polarizable FFs due to the more complex formulation and more parameters. Several tools are available for the parameterization of non-polarizable FFs,[202–204] while tools for polarizable FFs are still lacking. FFParam[205] provides the parameterization for Drude model, however, parameterization tools for CTPOL model are not yet available. **Paper III** implements CTPOL model on OpenMM and introduces the open source parameterization tool FFAFFURR, which supports the parameterization of OPLS-AA and

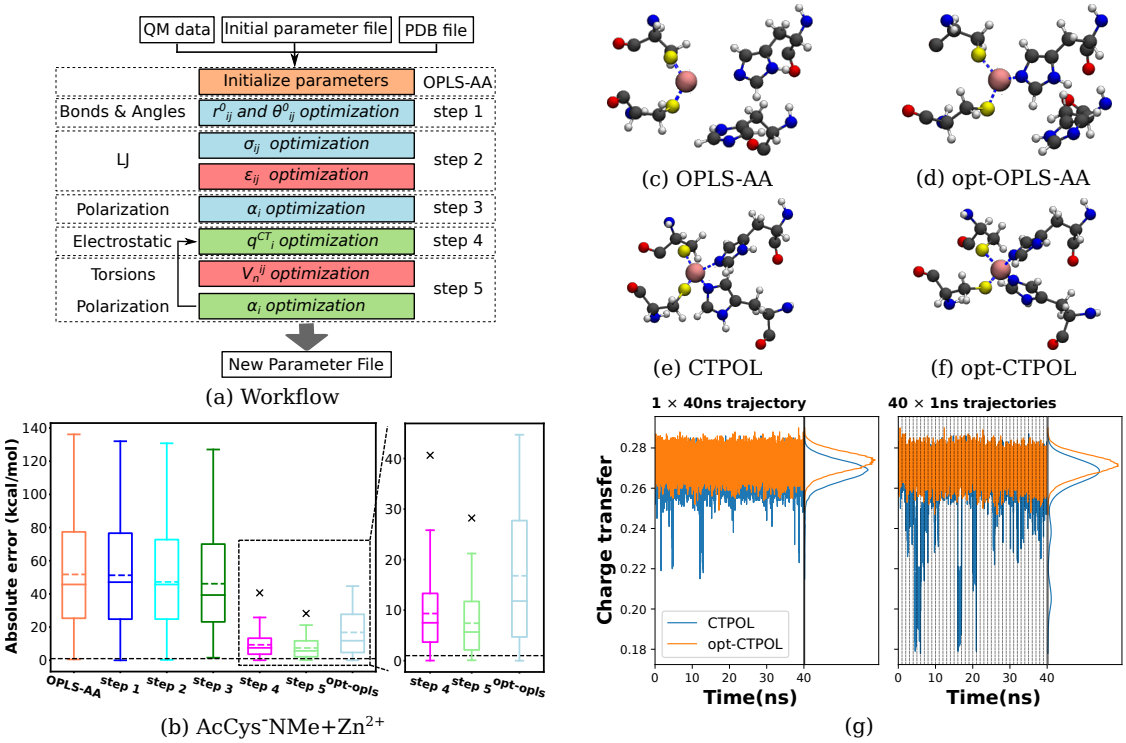


Figure 6: (a) Workflow of CTPOL parameterization. Parameters in blue boxes are derived from DFT calculations, parameters in coral boxes are obtained by Lasso or Ridge regression, and parameters in green boxes are obtained by PSO. (b) The absolute error distribution after each step in the workflow of AcCys$^-$NMe-Zn^{2+}, and the absolute error of OPLS-AA with fully optimized parameters (opt-opls) (c-f) The structures of interaction center of zinc finger protein after 40 ns MD simulation with different parameter sets. (g) Charge transfer along time with parameter sets CTPOL and opt-CTPOL.

CTPOL model. Since it has been found that both the Drude model and the CTPOL model require LJ parameterization in **Paper II**,[83] FFAFFURR provides the parameterization of all energy terms in the OPLS-AA and CTPOL models. Users can choose which energy term to adjust according to their needs in practical applications.

The performance of FFAFFURR was evaluated by its ability to reproduce relative energy hierarchies at the DFT level and by comparing the statistics of condensed-phase MD simulations. Figure 6 (a) and (b) display the workflow of CTPOL parameterization and the absolute error distribution after each step in the workflow of AcCys$^-$NMe-Zn^{2+}. Cysteine coordinated to Zn^{2+} serves as the center of many metalloproteins and has been reported to have significant charge transfer.[77] Figure 6 (b) shows that the absolute errors between FFs and DFT energies are significantly improved after the introduction of charge transfer for AcCys$^-$NMe-Zn^{2+}. After the full optimization of either parameter set, the CTPOL model better reproduces DFT energies than OPLS-AA.

Then, MD simulations of zinc finger protein (1ZNF) were used to validate the parameter set parameterized on the model peptides. Figure 6 (c) shows the structure of the interaction center of zinc finger protein after 40 ns MD simulation with OPLS-AA. The two histidines left the Zn^{2+}

center after the simulation. One potential usage of FFAFFURR is to optimize the parameters for the interaction center. The pair-wise LJ parameters of Zn^{2+} and atoms of histidine were optimized to obtain a new parameter set (opt-OPLS-AA). After the simulation, there was still one histidine leaving the interaction center as shown in Figure 6 (d). Then we tried to extend the opt-OPLS-AA model to the CTPOL model by introducing charge transfer and polarization term (opt-CTPOL) and succeeded in preserving the correct coordination number of the interaction center. To evaluate the effect of optimized pair-wise LJ parameters in the opt-CTPOL model, we tested the pure CTPOL model (CTPOL), which is the original OPLS-AA parameter set with charge transfer and polarization term introduced. It was found that the CTPOL model also preserved the correct coordination number of the interaction center. However, Figure 6 (g) shows the values of charge transfer along time with parameter sets CTPOL and opt-CTPOL. The data in the left panel are from a 40 ns MD simulation, and the data in the right panel are from 40 individual 1 ns simulations with different initial structures. The initial structures were derived from a short continuous MD simulation. The left panel shows that the charge transfer is more stable with opt-CTPOL, and the right panel further proves that the stability of the simulation using CTPOL is influenced by the initial structure. Overall, the results show that CTPOL model with the parameters derived from DFT data of model peptides can better simulate zinc finger proteins than classical FF, and the parameterization of LJ parameters is essential.

In conclusion, in the three papers that make up this thesis, I demonstrate:

- The creation of a comprehensive DFT data set for FF parameterization.
- The use of ontologies and FAIR data.
- Testing of classical and polarizable FF.
- Implementation of the CTPOL model and a parameterization tool.
- Testing of CTPOL model through MD simulations.

Chapter 4

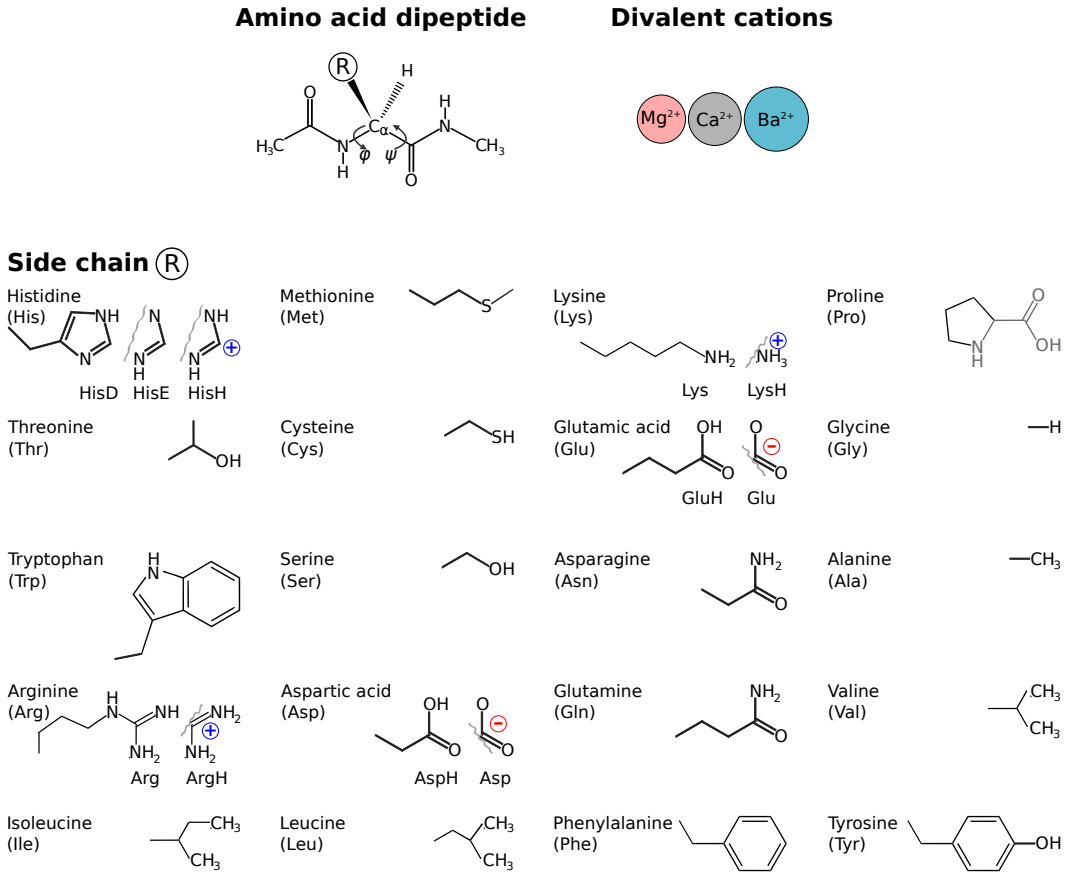
Publications

This Chapter presents the published scientific manuscripts which form the core of the thesis. The publications are organized thematically rather than chronologically. Before the manuscripts, an introductory page which represents publication title, authors' names, reference, URL, DOI and authors' contributions is provided.

4.1 Paper I: Better force fields start with better data: A data set of cation dipeptide interactions

X. Hu, M. O. Lenz-Himmer, C. Baldauf.

Sci. Data ***9****, 1-14 (2022)*



DOI: 10.1038/s41597-022-01297-3

Author contributions: I performed the calculations of all conformers, curated the data, constructed the ontology, and wrote the manuscript. M. O. Lenz-Himmer helped with the construction of ontology and contributed to the manuscript. C. Baldauf designed the study, curated the data, and wrote the manuscript.

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Better force fields start with better data: A data set of cation dipeptide interactions

Xiaojuan Hu ✉, Maja-Olivia Lenz-Himmer & Carsten Baldauf ✉

We present a data set from a first-principles study of amino-methylated and acetylated (capped) dipeptides of the 20 proteinogenic amino acids – including alternative possible side chain protonation states and their interactions with selected divalent cations (Ca^{2+}, Mg^{2+} and Ba^{2+}). The data covers 21,909 stationary points on the respective potential-energy surfaces in a wide relative energy range of up to 4 eV (390 kJ/mol). Relevant properties of interest, like partial charges, were derived for the conformers. The motivation was to provide a solid data basis for force field parameterization and further applications like machine learning or benchmarking. In particular the process of creating all this data on the same first-principles footing, i.e. density-functional theory calculations employing the generalized gradient approximation with a van der Waals correction, makes this data suitable for first principles data-driven force field development. To make the data accessible across domain borders and to machines, we formalized the metadata in an ontology.

Background & Summary

Metal cations are essential to life: one third of the proteins in the human body require metal cofactors[1,2]. By shaping the structure of proteins, cations affect biological processes like molecular recognition or enzyme activity. Understanding the structure, dynamics, and function of metalloproteins is in the ongoing focus of many researchers, we summarize a few examples that involve simulation approaches: Tamames *et al.* analyzed zinc coordination spheres in a data set from the Protein Data Bank and complemented with DFT-B3LYP calculations[3]. Sala *et al.* investigated folding of Pyrococcus furiosus rubredoxin (PfRd), which includes an iron ion, with classical molecular dynamics (MD) simulations[4]. A calcium binding site in the blood protein von Willebrand Factor (VWF) regulates force-triggered unfolding for cleavage and therewith its activity in primary hemostasis, as illustrated by classical force-probe MD simulations[5]. Gogoi *et al.* investigated protein-metal ion binding affinities by analysing MD simulations of 49 different cation-protein complexes[6]. Metal cations can alter peptide structure by interacting with backbones and thereby enforcing non-Ramachandran geometries[7]. Cations can, by repulsion or attraction, also substantially reduce the conformational flexibility of functional sidechains[8,9].

MD simulations of biomolecules typically rely on additive force fields, where distinct terms describe bonded and non-bonded interactions based on empirically derived parameters. Studies have shown that the accuracy of force fields is especially limited when describing interactions involving ionic species[10–13]. In particular non-bonded interactions are critical, but of course the effect that nearby located cations exert on bonds is almost impossible to grasp by the combination of bonded and non-bonded interactions in a general-purpose force field. Modeling of electrostatic interactions via pairwise Coulomb potentials is based on assigning partial charges to atoms[14]. Partial charges are derived by: (i) fitting to experimental data (GROMOS and OPLS prior 2005), e.g. by fitting partial charges to reproduce hydration free enthalpies[15,16], (ii) deriving partial charges from QM calculations (Amber and Charmm)[17,18], or the combination of the two strategies (OPLS after 2005)[19].

The reliability of a force field also depends on the physics behind the formulation. The failures of established biomolecular force fields when describing cation-peptide systems may result from a central underlying assumption – modeling atoms by fixed point charges and neglecting charge transfer and polarization effects, while both are crucial to ionic systems[20–23]. Introducing more physics to the model appears a promising route to improve force fields: The inclusion of electronic polarization and charge transfer plays a central role in the next generations of biomolecular force fields[24–26]. However, including additional terms leads to force fields with way more

Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195, Berlin, Germany. ✉e-mail: xhu@fhi.mpg.de; baldauf@fhi.mpg.de

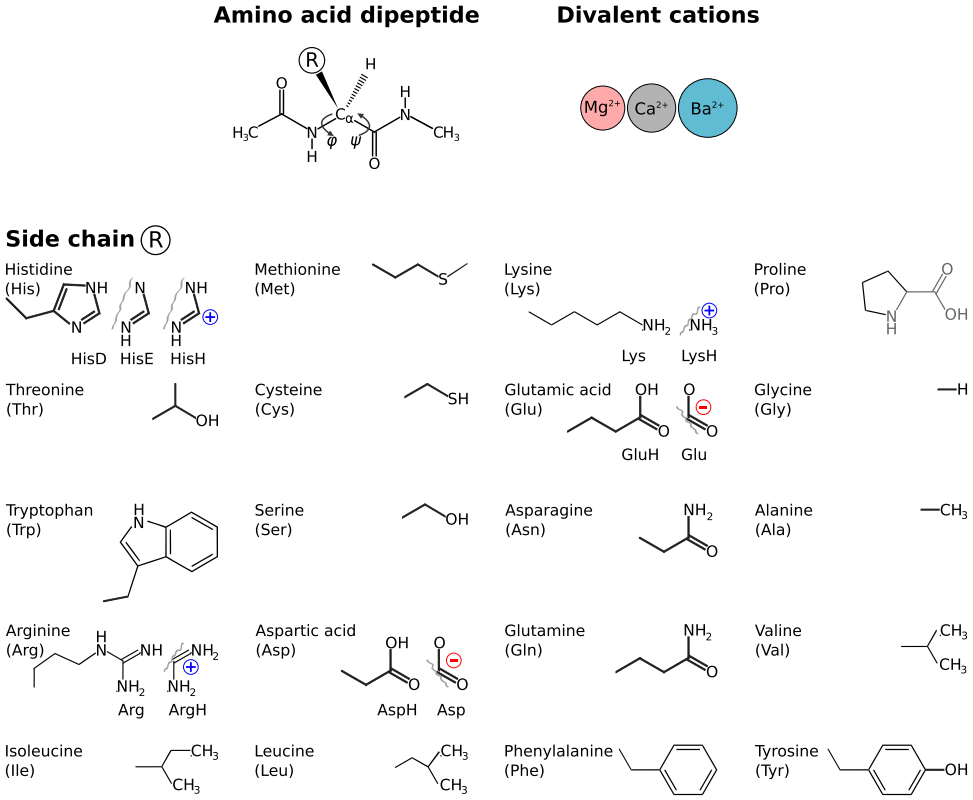


Fig. 1 The molecular systems in this study are dipeptides of the 19 proteinogenic amino acids that differ in the side chain **R** and the proteinogenic imino acid proline. Where applicable, different protonation states were considered.

parameters, which makes parameterization more challenging[27,28], in particular in the absence of high-resolution experimental data of less stable conformations, i.e. higher-energy structures[29]. To summarize, we see three main challenges:

- The availability of sufficiently-accurate electronic-structure data as well as choosing the "right ways" to derive e.g. partial charges from it.
- Designing the formulation of next-generation force fields that also include, for example, charge transfer and polarization.
- Finding sets of parameters (force fields) for such potentials in the absence of experimental data at sufficient spatial and time resolution.

Thorough studies have deepened our understanding of the conformational basics of individual building blocks, e.g.[30–41]. However, these studies are highly diverse with regards to the approximations made to model and to search the potential energy surfaces (PES) of the respective molecular systems; furthermore, the data is often not available. The availability of uniform and comprehensive computational data at an appropriately accurate level of theory has the potential to substantially increase the predictive power of force fields[42]. In order to provide such amino acid data sets for force field development on consistent computational footing, we extend previous work[43] by focusing on dipeptides as models of amino acid building blocks in polypeptide chains in complex with the divalent cations Mg^{2+}, Ca^{2+}, and Ba^{2+}, which play prominent roles in physiology: Mg^{2+} takes structural, catalytic, and regulatory roles[44] regulating ion channels, mitochondrial function, and cell's pH and volume[45]. Ca^{2+} levels regulate muscle contraction, hormone secretion, metabolism, ion transport, division, *etc*[46]. Mg^{2+} and Ca^{2+} may compete for the same binding sites[47]. Ba^{2+} can cause cardiac irregularities and affect the nervous system presumably by blocking potassium channels[48].

Combining these 3 cations with the proteinogenic amino acids in all meaningful side chain protonation states results in a data set that covers a wide range of molecular systems, see Fig. 1.

For the 21,909 stationary points, properties relevant to force field development were computed, details can be found in the Methods section. Making the data FAIR[49,50] – as in findable, accessible, interoperable, and reusable – is a challenge. In particular as we want to make the data available also to experts from other domains of science or to autonomous agents. To that end, we make the data freely available and also provide ontologies. An ontology defines a common vocabulary of basic concepts in a domain and relations among them[51]. The benefit is primarily that these definitions are machine-readable. This allows for interoperability between resources and databases as well as data interpretation across data collections. Through developed ontological representation of the data set, it can be connected to upper level concepts and thereby made machine-usable, which in turn

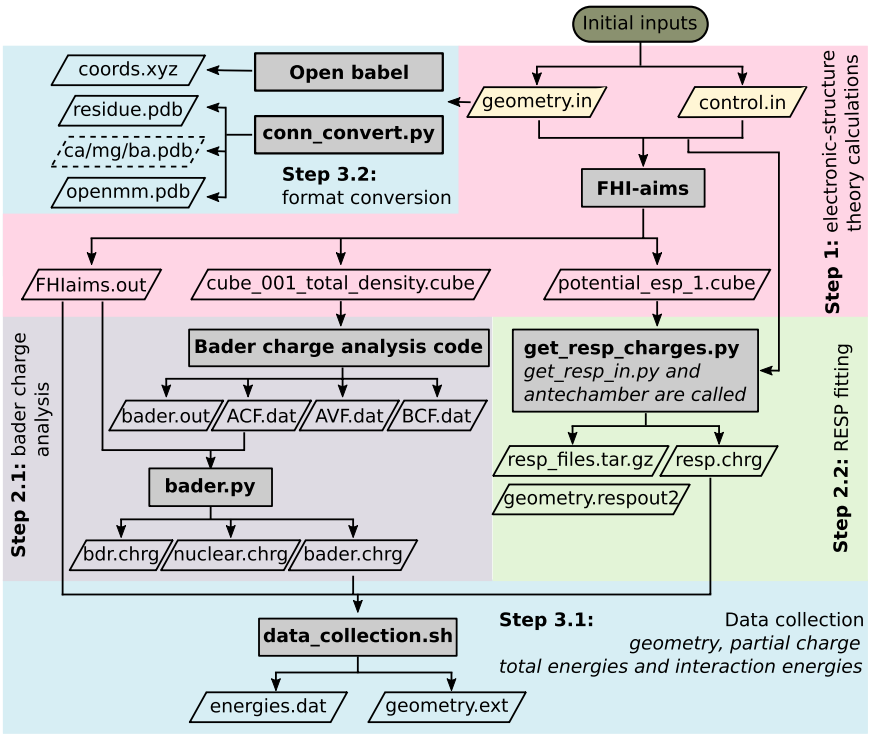


Fig. 2 Schematic representation of the workflow employed to derive properties of each conformer. Calculation steps were displayed in boxes with different background colour. Gray boxes indicate tools employed in each step. Parallelograms represent input and output files in each step. Links to custom codes are listed in Section *Code availability*.

enables automatic access and querying of the data. Ultimately, researchers can share their data with experts from other domains as well as making data available to machine intelligence.

Methods

Figure 1 summarizes the molecular systems in this study. Including the protonation states, we have to consider 26 dipeptides in 4 complexation states (bare, Ca^{2+}, Mg^{2+}, Ba^{2+}) which results in the 104 systems for which our structure searches identified 21,909 stationary points. For each of these stationary points, not only structure and energy are provided, but also further properties relevant to force field development, namely: van der Waals energies, interaction energies as well as electron densities and derived properties like the electrostatic potential, diverse partial charge models, and effective atomic volumes. By that, our dipeptide-cation data set allows one to explicitly assess subtle, but important, effects of local changes in the electrostatic environment due to peptide-cation interaction.

Sampling method. A hierarchical structure search that is described in detail in reference[43] was employed to locate stationary points on the potential energy surfaces of the 104 molecular systems. The initial global conformational searches of all dipeptides with/without Ca^{2+} were performed by a basin hopping search strategy[52,53] using the OPLS-AA force field[16]. Secondly, a refinement using density-functional theory calculations was performed. All electronic-structure calculations were performed with the all-electron, full-potential code FHI-aims utilizing numeric atom-centered basis functions[54–56]. The PBE generalized-gradient exchange-correlation functional[57] augmented by Tkatchenko's and Scheffler's pairwise van der Waals correction[58] was employed, and is referred to as PBE+vdW throughout this work. Stationary points that resulted from the FF-based pre-sampling were subjected to DFT-PBE+vdW relaxations with `light` settings. Next, a local first-principles based sampling step by *ab initio* replica-exchange molecular dynamics (REMD)[59,60] employing DFT-PBE+vdW with `light` settings, was applied to the identified set of structures. Conformers were extracted every 10 steps from REMD trajectories and clustered with a *k*-means clustering algorithm[61]. Obtained conformers went through relaxation with PBE+vdW (`light` computational settings), clustering and further relaxation with PBE+vdW (`tight` computational settings) to obtain the final conformational hierarchies. Initial structures of Mg^{2+} and Ba^{2+} binding dipeptides were obtained by substituting Ca^{2+} cation in dipeptide binding a Ca^{2+} cation. Subsequently, those were put into the procedure from *ab initio* REMD simulations to relaxation with PBE+vdW (`light` computational settings) to obtain final conformers as described before. These structures were further relaxed by PBE+vdW with `tight` computational settings.

Property calculations. Property calculations were performed on all structures obtained by the sampling method described above. This includes also high energy conformers. Figure 2 shows the processes involved in the property calculations; the individual steps are described in detail below. From the PBE+vdW DFT calculations with `tight` computational settings using FHI-aims, we collect in **Step 1** total energies, vdW energies, interaction energies, electron densities, electrostatic potential, Hirshfeld partial charges[62], and effective atomic volumes. Based on the effective atomic volumes V^{eff} per atom we provide, the effective vdW radii (R^0_{eff}) and the polarizability (α^0_{eff}) of an atom in a molecule can be calculated as follows[58,63]:

$$R^0_{\text{eff}} = R^0_{\text{free}} \left(\frac{V^{\text{eff}}}{V^{\text{free}}} \right)^{1/3} \tag{1}$$

$$\alpha^0_{\text{eff}} = \alpha^0_{\text{free}} \left(\frac{V^{\text{eff}}}{V^{\text{free}}} \right) \tag{2}$$

$$\frac{V_i^{\text{eff}}}{V_i^{\text{free}}} = \frac{\int r^3 \omega_i(\vec{r}) n(\vec{r}) d^3\vec{r}}{\int r^3 n_i^{\text{free}}(\vec{r}) d^3\vec{r}} \tag{3}$$

in which, R^0_{free} and α^0_{free} are the vdW radii of reference free-atom and static dipole polarizability (which can be taken from either experimental data or high-level quantum chemical calculations), respectively. V^{free} is the volume of the free atom *in vacuo*, r^3 is the cube of the distance from the nucleus of atom i, $\omega_i(\vec{r})$ is the Hirshfeld atomic partitioning weight for atom i, $n(\vec{r})$ is the total electron density, and $n_i^{\text{free}}(\vec{r})$ is the electron density of the free atom i.

The basic property resulting from a DFT calculation is the electron density, which – for each entry in our data set – was stored on a discrete grid of points with a spacing of 0.05 Å in a rectangular volume, which spans the whole molecule plus 14 Bohr (7.4 Å) beyond the outermost nuclei. The electrostatic potential exerted by a molecule on its environment may be used to derive partial charges. To that end, for each entry in the data set, five molecular surfaces were created by increasing the van der Waals radii of all atoms in the molecule (molecule with cation) by factors between 1.4 and 2.0. Points on these surfaces were represented in a cubic grid of each 35 grid points in x, y, and z direction. For these points, the electrostatic potential was evaluated. For biomolecular force fields, atomic partial charges are a crucial ingredient for computing the pairwise Coulomb term of the non-bonded interactions. We provide three types of partial charges:

- Hirshfeld atomic charges, computed by FHI-aims, were derived based on the Hirshfeld partitioning scheme[58,62]. The Hirshfeld atomic charge q_i of atom i is given by

$$q_i = Z_i - \int n_i(\vec{r}) d^3\vec{r} \tag{4}$$

where Z_i refers to the corresponding atomic number, and $n_i(\vec{r})$ is the associated electron density associated with atom i.

$$n_i(\vec{r}) = \omega_i(\vec{r}) n(\vec{r}) \tag{5}$$

where $n(\vec{r})$ denotes the total electron density, $\omega_i(\vec{r})$ is the Hirshfeld atomic partitioning weight for atom i. $\omega_i(\vec{r})$ is given by

$$\omega_i(\vec{r}) = \frac{n_i^{\text{free}}(\vec{r})}{\sum_A^{Allatoms} n_A^{\text{free}}(\vec{r})} \tag{6}$$

- Bader charges were being computed in **Step 2.1** using the Bader Charge Analysis tools[64–66] provided by the Henkelman group based on the electron density cube file produced in Step 1. The atoms in molecules (AIM) partitioning method uses what is called zero flux surfaces to distribute electron density among the atoms. Such zero flux surface is a two-dimensional surface on which the charge density is a minimum perpendicular to the surface. In molecular systems, the charge density typically reaches a minimum somewhere between pairs of neighboring nuclei. This can be seen as the natural place to separate atoms from each other. These borders between atoms define the electron density region associated with a given atom, from which the partial charges are being calculated.
- In **Step 2.2**, RESP partial charges[67–69] were computed using Antechamber[70] from the AmberTools package[71]. A two-stage restrained electrostatic potential (RESP) fitting procedure[67] was employed as implemented in Antechamber.

In the final **Steps 3.1 and 3.2**, data was collected and files converted to established formats. Geometry information is provided in three formats: the FHI-aims input format, the xyz format generated by Open Babel[72],

and PDB files that are readable by the CHARMM-GUI portal[73] and the openMM7 package[74]. Connectivity and atom type information – needed for the PDB format – was gathered based on atomic distances by the Python script `conn_convert.py`. Furthermore, energies and partial charges were tabulated for convenient usage. Interaction energies E_{inter} between cation and dipeptide were calculated as follows:

$$E_{\text{inter}} = E_{\text{complex}} - E_{\text{dipeptide}} - E_{\text{cation}} \tag{7}$$

where $E_{complex}$ corresponds to the potential energy of the dipeptide-cation complex, $E_{dipeptide}$ is the potential energy of the dipeptide alone fixed in the cation bound conformation, and E_{cation} is the potential energy of the isolated cation.

Further data and properties can be extracted from the raw and normalized data[75] that is available from the NOMAD Repository and Archive[76]. The data set was deposited as populated ontology in OWL format[77] in the EDMOND repository of the Max Planck Society. The construction of the ontology is described in the following subsection.

Ontology construction. Ontology construction is an iterative process involving many steps from defining common vocabularies, identifying the most important concepts and their relations to modelling such concepts in a semantically correct and still useful and applicable way. It can be used to enrich, annotate, and link data that is then called *linked data* and usually expressed in a semantic triple format consisting of *subject*, *predicate*, and *object*[78]. The main components of an ontology are classes, properties, individuals and axioms. Classes are the focus of most ontologies and are descriptions of concepts in a domain and represent a specific set of individuals. "Ala" is a class in the Amino Acid domain, thus each single Ala conformer in our data set is an individual of class "Ala". Properties describe features and attributes of classes and individuals. Properties can connect classes and individuals. For example, *hasProperty* can connect classes "Ala" and "Charge" as a property. Axioms are statements that all together define what is the truth in a given domain. In this work, the ontology builder Protégé[79] and the python package Owlready2[80] were employed to build ontologies in the OWL2 Web Ontology Language (http://www.w3.org/TR/owl2-overview) which is based on RDF – the Resource Description Framework (http://www.w3.org/TR/rdf-primer). Subjects and predicates are named using Internationalized Resource Identifiers (IRIs) (https://tools.ietf.org/html/rfc3987), while the object position can be filled by an IRI or a literal value (e.g. string or number). Ontologies created in this work have been tested with the OWL reasoner FACT++[81].

Data Records

Raw data and normalized data of the DFT calculations for this amino acid dipeptide data set is available from the NOMAD repository (http://nomad-repository.eu) via the https://doi.org/10.17172/NOMAD/2021.02.10-1[75]. The NOMAD Archive contains all raw input, output, and property calculation files for download, while the NOMAD Repository contains normalized data, i.e. a digest of the DFT calculations. Data in the NOMAD Repository and Archive is provided on the basis of the Creative Commons Attribution 3.0 License (CC BY 3.0) as it is stated in the NOMAD terms (https://nomad-lab.eu/terms).

The extracted data in form of a populated ontology in OWL format is available download via the https://doi.org/10.17617/3.5q10.17617/3.5q[77] under the Creative Commons Attribution 4.0 license (CC BY 4.0). In the following two subsections, we briefly introduce the data and the concept of the provided ontology.

DFT data set. The distribution of the 21,909 stationary points of the amino acid dipeptide (plus cation) systems over the different amino acid building blocks is summarized in Fig. 3. This data is in particular intended for training energy functions in machine learning approaches in the context of force field development and parameterization. Consequently, it consists not only of geometries with total energies for preferred low-energy conformers. Instead, DFT-PBE+vdW calculations also included high-energy conformers. The data we provide is particularly focused on parameterizing non-bonded interactions: The above-mentioned cation-peptide interaction energies were already used to tune force fields parameters of non-bonded interactions[26,82]. The comparison to DFT-based vdW energies computed with the Tkatchenko-Scheffler formalism[58] is useful to evaluate or adjust the non-bonded Lennard-Jones parameters ε and σ. Importantly, due to the spread over high and low energy conformations, diverse substructures and environments (due to cation binding), a range of partial charge values is sampled that informs about polarization and charge transfer. To that end, the electronic structure is simplified into partial charge models, based on Hirshfeld partitioning or Bader AIM analysis of the electron density. The electron density, in combination with the nuclear charges, also defines the electrostatic potential (ESP) around the molecule, which can be used to derive force field parameters related to electrostatic interaction[83]. The electron density has been used before to derive environment-specific force fields[84]. Electron densities for a large set of molecules have been used to predict partial charges based on machine learning[85,86], to that end, an average over similar substructures in different molecules was used.

The data is first of all made available as a set of files. The different files, their content, and which programs to read or write them are given in Table 1. A direct way to access the data is to download the compressed archive[75] and browse the folder structure that is given in Fig. 4 or download from the same source the normalized data in json-files.

This way of representing data however limits the automated access to the data by artificial agents or by researchers from other domain, as the metadata to the data is somewhat hidden. In order to alleviate this, the next section details the ontology which we developed in order to provide an extensible, machine-interpretable and machine-usable model for the automated access and post-processing of the data set.

Ontology. AAMI (Amino Acid Meta-Info) is an ontology created "bottom-up" to specifically represent the meta-information of this amino acid-cation data set in a machine-understandable and machine-processable way.

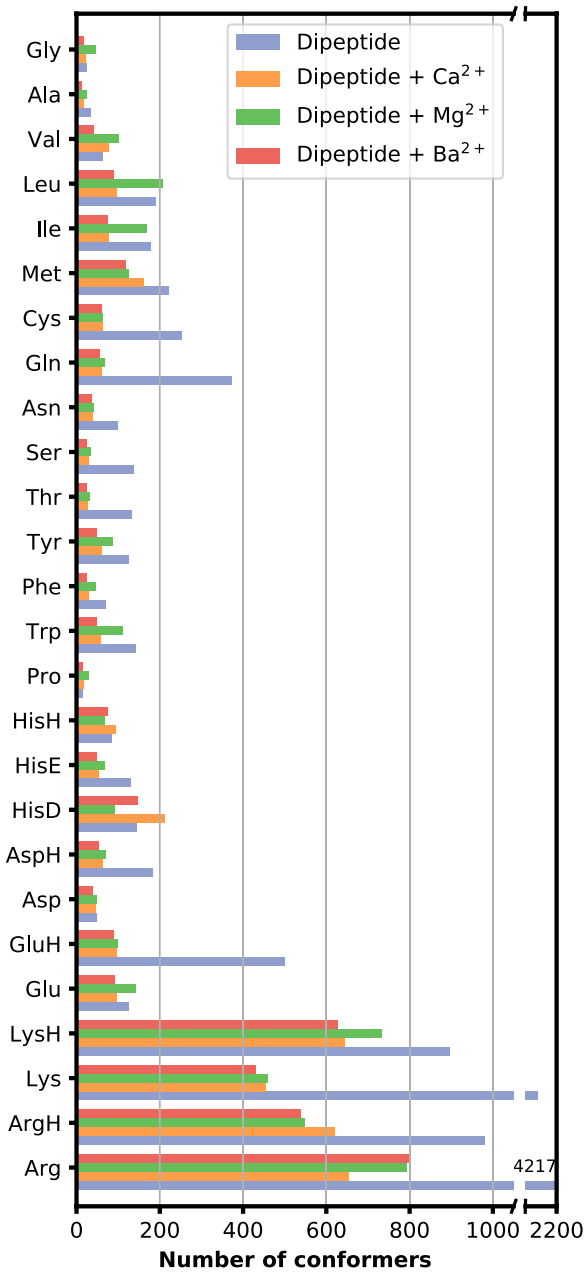


Fig. 3 Numbers of stationary points of each molecular system covered in this study.

AAMI does not only contain metadata of properties, it also covers processes of analysis, such as inputs, outputs, and tools in each process and their roles, which further makes data interpretable and understandable. Two existing ontologies were re-used in AAMI: the European Materials Modelling Ontology (EMMO) (https://emmc.info/emmo-info), which provides a representational framework for materials modelling and characterization knowledge, and the Amino Acid Ontology (http://bioportal.bioontology.org/ontologies/AMINO-ACID), which provides structured knowledge of amino acids and their properties. By reusing existing terms in EMMO and Amino Acid Ontology rather than creating the ontology from scratch, terms in AAMI were connected to upper level concepts and can be potentially linked to further ontologies. Moreover, users are able to take advantage of data and annotations that are already used in those ontologies and can by that also rely on concepts that were already agreed upon in a bigger community. The primary aim of AAMI is to make our data set FAIR (Findable, Accessible, Interoperable, and Reusable)[49], in particular accessible, interoperable and reusable. The elements of AAMI can be found in Fig. 5. In the AAMI ecosystem, we created:

File name	Description	Code/Format
FHI-aims Input Files		
geometry.in	Cartesian coordinates of the complexes	FHI-aims
control.in	Input file with technical parameters for electronic structure calculations	FHI-aims
FHI-aims Output Files		
FHIaims.out	Main output the electronic structure calculations, contains: total energy, vdW energy and effective atomic volume etc.	FHI-aims
cube_001_total_density.cube.bz2	Cube file representation of the electron density (bzip2 compressed)	FHI-aims
potential_esp_1.cube.bz2	Cube file representation of the electrostatic potential (bzip2 compressed)	FHI-aims
hirsh.chrg	Hirshfeld charges	Self-made
Geometries		
coords.xyz	Coordinate file	*xyz* format
residue.pdb	Coordinate file	CHARMM
[cation].pdb	Separate coordinate file for each of the cations Ca, Ba, Mg	CHARMM
openmm.pdb	Coordinate file	OpenMM
Bader AIM calculations		
ACF.dat, AVF.dat, BCF.dat, bader.out	Information of Bader charge analysis	Bader
nuclear.chrg, bdr.chrg	Information of Bader charge analysis	Self-made
bader.chrg	Bader charges	Self-made
RESP calculations		
geometry.respout2, resp_files.tar.gz	RESP charge information	Antechamber
resp.chrg	RESP charges	Self-made
Aggregated output		
geometry.ext	Collection of coordinate and charge information	Self-made
energies.dat	Collection of total energy and interaction energy	Self-made

Table 1. List and description of file types in the data set.

1. The cluster structure ontology (CSO) represents concepts and relations for structure description of non-periodic systems, EMMO was imported, and 351 classes and 2053 axioms were created.
2. The cluster property ontology (CPO) describes properties of non-periodic systems. CSO was imported, and 450 classes and 2984 axioms were created.
3. The force field ontology (FFO) represents concepts in force fields, e.g. atom type and atom class. Amino acid ontology and CPO were imported, and 563 classes and 4453 axioms were created.
4. AAMI represents concepts and relations in the amino acids-cation data set. FFO was imported, and 787 classes and 5466 axioms were created.
5. The different instances of AAMI-D-* are knowledge graphs created from the data set in this study. Such graph is build by populating AAMI with the data for an amino acid, e.g. ALA, ARG, *etc*., from this data set. The asterisk represents the name of the corresponding amino acid.

Partial high level class organization and some of the classes and relations of AAMI are shown in Fig. 6 to give an overview of the organization of the ontology and how terms from each ontology are related to each other.

The primary use of AAMI is to annotate database records. However, since ontologies were developed with the OWL2 Web Ontology Language, which represents data by sets of subject-predicate-object statements, so-called *triples*, the underlying computational logic enables automatic inference and querying over data repositories. In principle, any question framed in the respective mathematical logic can be answered in a finite number of steps. However, such reasoning capabilities are currently limited to description logic. Data query can be done with the ontology and linked data query language, SPARQL (https://www.w3.org/TR/sparql11-query). A user can query for sub-classes, relations between classes, functional annotation, and so on. Stardog Studio (https://www.stardog.com/studio) can be used as a *triple store* and employed to perform the SPARQL queries. A tutorial of SPARQL query language using Stardog Studio can be found in the following link: https://www.stardog.com/tutorials/sparql/. We provide two sample queries in this work to guide users to build their own queries.

Before any queries, a set of namespace prefixes were declared to abbreviate IRIs, e.g. the knowledge graph of alanine dipeptide was imported as an example under the `PREFIX ala`.

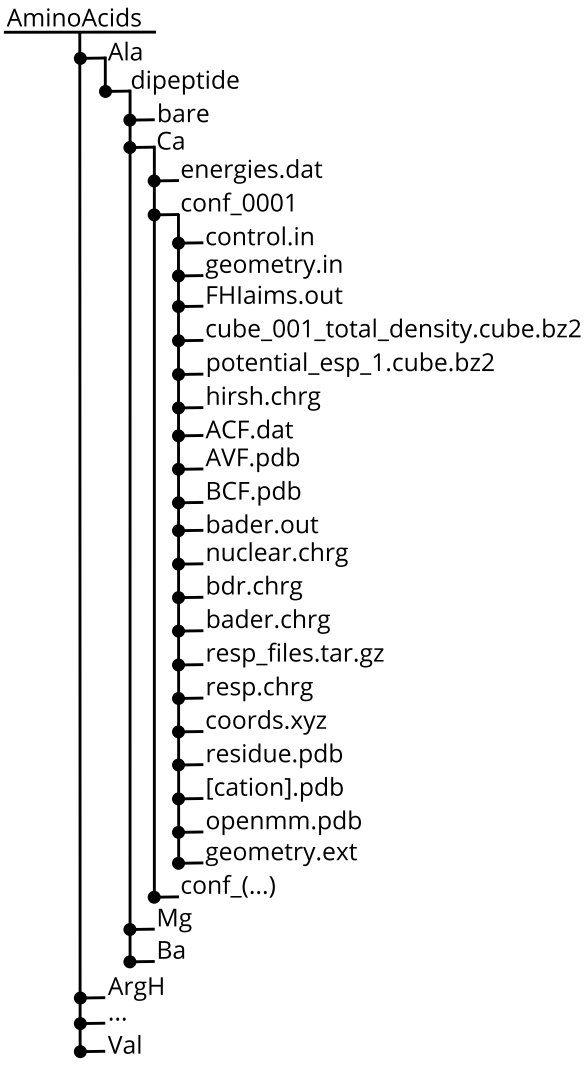


Fig. 4 Schematic representation of the folder structure of the data. Each folder, as exemplified for the Ca^{2+}-coordinated cysteine dipeptide, contains multiple properties per system.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX cso: <http://www.semanticweb.org/ClusterStructure.owl#>
PREFIX cpo: <http://www.semanticweb.org/ClusterProperty.owl#>
PREFIX ffo: <http://www.semanticweb.org/ForceField.owl#>
PREFIX aami: <http://www.semanticweb.org/AAMI.owl#>
PREFIX ala: <http://www.semanticweb.org/AAMI-D-Ala-Dipeptide.owl#>
PREFIX hasProperty:<http://emmo.info/emmo/middle/properties#
    EMMO_e1097637_70d2_4895_973f_2396f04fa204>
PREFIX hasSymbolData:<http://emmo.info/emmo/middle/perceptual#
    EMMO_23b579e1_8088_45b5_9975_064014026c42>
```

The main query form in SPARQL is a `SELECT` query. A `SELECT` query has two main components: a list of selected variables and a `WHERE` clause for specifying the graph patterns to match. For example, according to the graph shown in Fig. 6, we can query for Bader charges of atoms which have atom type of "1" in Amber10 with a `SELECT` query as follows:

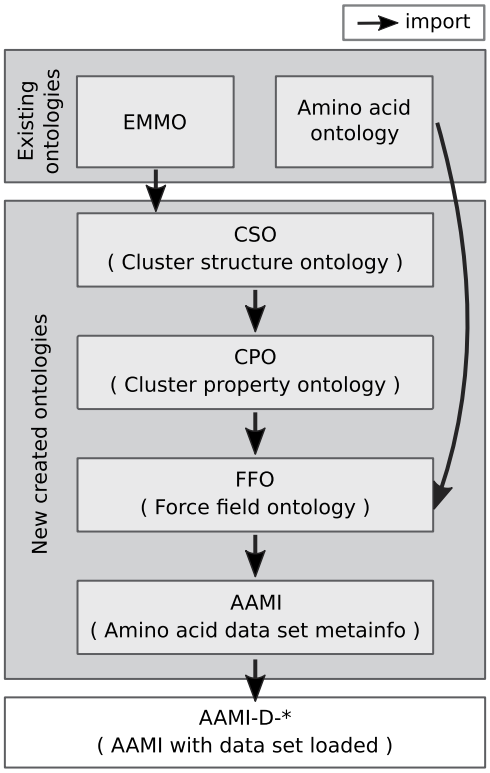


Fig. 5 Hierarchy of the ontologies linked to amino acid-cation meta-info (AAMI). Details of the ontologies and relations among them are described in Section *Ontology*.

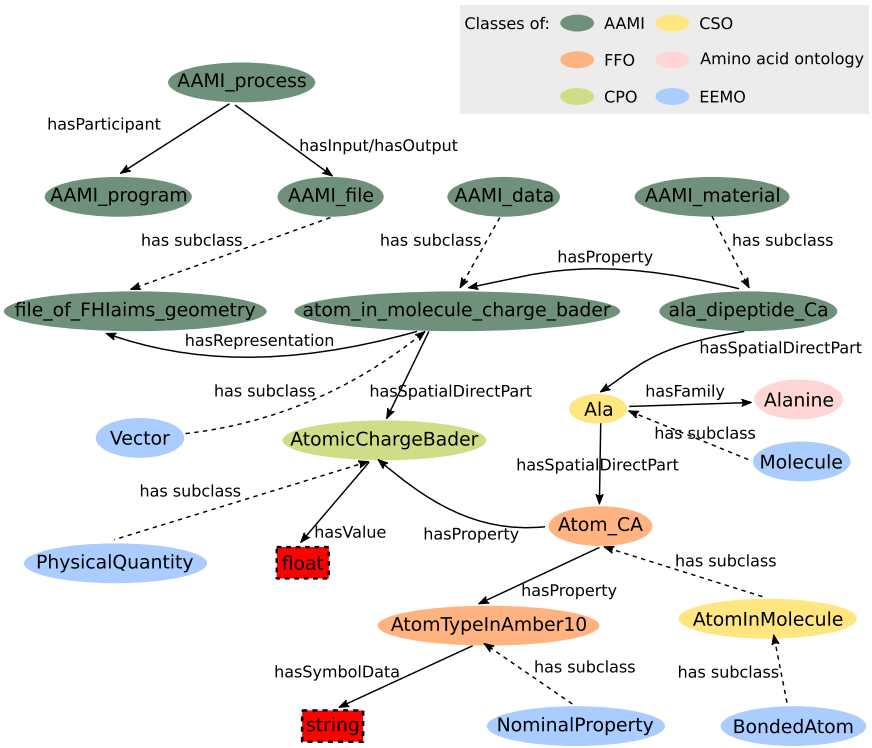


Fig. 6 Partial high-level class structure of AAMI ontology. Ovals represent classes, where classes from different ontologies are color coded. Rectangles represent literals. Solid lines are properties and dotted lines represent the relation of '*has subclass*'.

```
[...]
SELECT ?atom ?n
WHERE {
   ?atom hasProperty: ?atomtype.
   ?atomtype a ffo:AtomTypeInAmber10.
   ?atomtype hasSymbolData: "1"^^xsd:string.
   ?atom hasProperty: ?badercharge.
   ?badercharge a cpo:AtomicChargeBader.
   ?badercharge cpo:hasValue ?n
}
```

The resulting list shows all atoms of type "1" in Amber10, *i.e.* hydrogen atoms bound to a peptide bond nitrogen, and their Bader charges:

```
ala#Atom_HN_11_alaD_Ca_conf_0017 0.450512
ala#Atom_HN_11_alaD_Ca_conf_0018 0.486539
ala#Atom_HN_11_alaD_Ca_conf_0014 0.450169
ala#Atom_HN_11_alaD_Ca_conf_0012 0.484383
ala#Atom_HN_11_alaD_Ca_conf_0002 0.442222
ala#Atom_HN_11_alaD_Ca_conf_0006 0.452150
...
```

Another useful query is `DESCRIBE`, which returns all the outgoing edges of a node. `DESCRIBE` is most useful when we don't know much about the ontology and want to quickly see the terms used in the triples. For example, we can query "describe individuals which belong to class Atom_C" with `DESCRIBE` query within the alanine dipeptide knowledge graph:

```
[...]
DESCRIBE ?atom
WHERE {
   ?atom a ffo:Atom_C
}
```

In the following, we display part of the output of the query, from which we can see that an individual "Atom_C_9_alaD_Ca_conf_0017" belongs to class "Atom_C" and has properties of "AtomicChargeBader_1.35427", "position9" and so on.

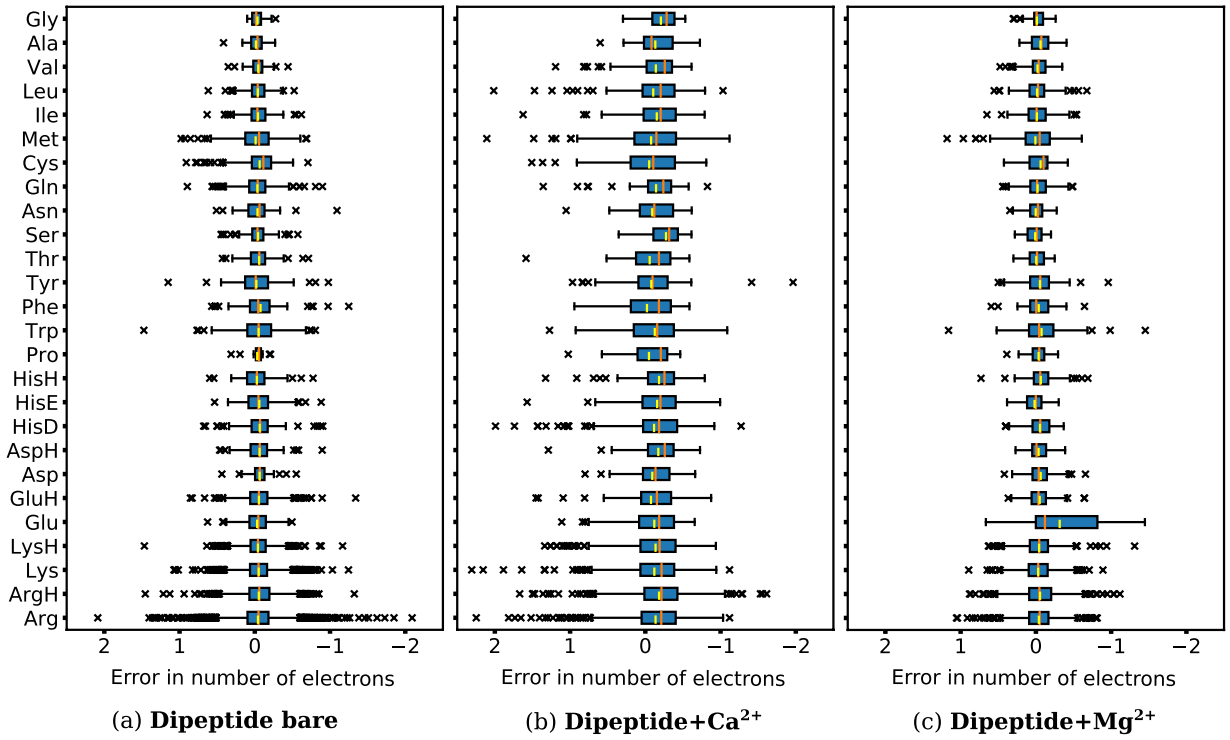
```
@prefix ffo: <http://www.semanticweb.org/ForceField.owl#> .
@prefix ala: <http://www.semanticweb.org/AAMI-D-Ala-Dipeptide.owl#> .
@prefix hasProperty: <http://emmo.info/emmo/middle/properties#
    EMMO_e1097637_70d2_4895_973f_2396f04fa204> .

{
   <ala#Atom_C_9_alaD_Ca_conf_0017> a owl:NamedIndividual , ffo:Atom_C ;
     <hasProperty> <ala#AtomicChargeBader_1.35427> , <ala#EffectivePolarizability_9
         .839967844590108> , <ala#position9>, ...
}
```

With tools like Stardog Studio, the results of such query can be written out in various file formats for further usage, e.g. `XML`, `JSON-LD` for triples output or `CSV` for tabular output.

Technical Validation

The reliability of the DFT-PBE + vdW level of theory for amino acids and amino acids binding divalent cations was evaluated before[43]. In this reference, single-point energy calculations were performed on all structures of alanine (Ala) and phenylalanine (Phe) amino acids in isolation, as well as binding with a Ca^{2+} cation employing Møller-Plesset second-order perturbation theory (MP2)[87,88]. For the structures of the amino acids Ala and Phe without cation bound, mean absolute errors (MAE) within chemical accuracy (1 kcal/mol) were estimated for PBE + vdW. A different long-range dispersion method, the many-body dispersion model (PBE + MBD)[89], didn't show significant improvements for isolated amino acids. Also the usage of a hybrid exchange-correlation functional, PBE0 (PBE0 + MBD)[89], did not significantly improve the MAEs. However, the maximum error of Phe was reduced from 2 kcal/mol to 1.3 kcal/mol. MAEs were slightly higher with PBE + vdW when Ca^{2+} was involved. They reached 1 kcal/mol and 2 kcal/mol for Ala + Ca^{2+} and Phe + Ca^{2+}, respectively. Employing both, many-body dispersion and the hybrid functional PBE0, improved the MAE to about 1 kcal/mol. In a manuscript on histidine-zinc interactions[11], DLPNO-CCSD(T)[90,91] was employed to benchmark several DFAs as well as the wave function-based MP2 method. The evaluated systems are (a) negatively charged acetylhistidine (AcH) with and without a Zn^{2+} cation, and (b) neutral AcH with and without a Zn^{2+} cation. The results showed that PBE+vdW gave an acceptable accuracy. In conclusion, PBE+vdW appears to be a valid starting point for studies on cation-peptide systems.

The validation of the sampling method can be elucidated by the work in ref. [92]. A genetic algorithm was employed to do the sampling of the low-energy segment in the conformational space of seven dipeptides:



(a) **Dipeptide bare** (b) **Dipeptide+Ca^{2+}** (c) **Dipeptide+Mg^{2+}**

Fig. 7 Error in numbers of electrons from Bader analysis of Dipeptide (**a**) bare, (**b**) with Ca^{2+} and (**c**) with Mg^{2+}. The upper and lower lines of the rectangles mark the 75% and 25% percentiles of the distribution, the orange and yellow horizontal lines in the box indicate the median (50% percentile) and mean value, and the upper and lower lines of the "error bars" depict the 99% and 1% percentiles. Crosses represent the outliers.

Glycine (Gly), Alanine (Ala), Phenylalanine (Phe), Valine (Val), Tryptophan (Trp), Leucine (Leu), Isoleucine (Ile). Conformers from our previous data set[43] were used as reference points and both studies agree in their overall structure findings.

The potential usage of our data set has been confirmed in ref. [26]. In this work, our data set was used to assess the accuracy of existing FFs by their abilities to reproduce quantum mechanical (QM) interaction energies of Ca^{2+}-dipeptide. By relating the parameter space to conformational space, the utility of our data set as a reference for future optimization of polarizable force fields is illustrated.

An assessment of the reliability of Bader charge analysis of bare dipeptides as well as dipeptide-Ca^{2+} and dipeptide-Mg^{2+} complexes is shown in Fig. 7. The number of electrons from Bader charge analysis yielded high errors in some structures of dipeptide-Ca^{2+}, reaching 2 electrons. This error apparently results from too wide grid spacing at regions of rapid density change (near "heavy" cores) when writing the electron density to cube files, the input for the Bader analysis code. Changes in electron density are particularly large close to the cations in the investigated clusters, so in principle grid spacings adjusted to the respective systems would be required. Overall, however, the mean errors of each amino acid are around 0. The errors of dipeptide-Mg^{2+} have the same trend, but are smaller than the errors of dipeptide-Ca^{2+} due to the smaller radius of Mg^{2+}. Ba^{2+} is much heavier than Ca^{2+} and Mg^{2+}, the rise in density close to the atomic center is much steeper. To analyze the Bader charges of dipeptide-Ba^{2+} complexes, a much smaller grid spacing is needed. However, this will result in electron density cube files that are impractically large for an overview study of this extend. So in this work, we did not present the electron density and Bader charges of dipeptide-Ba^{2+} complexes.

Usage Notes

Attention, the download of the whole archive of raw data is about 1.5 TB in size (compressed). Structures in this data set are stationary-point geometries, most of them can be expected to be minima, yet there are certainly also saddle points. All files in the NOMAD repository can be downloaded through `curl` based on upload and entry IDs (variables: `upload_id` and `entry_id` below). The command below downloads all files in one calculation:

```
curl "http://repository.nomad-coe.eu/app/api/raw/calc/upload_id/entry_id/*" -o download.zip
```

The metadata for the DFT calculations can in part be browsed at the NOMAD Archive page (https://www.nomad-coe.eu/the-project/nomad-archive/archive-meta-info). There are numerous tools to perform SPARQL queries, e.g. Stardog Studio (https://www.stardog.com/studio), Protégé[79], RDFLib (https://github.com/RDFLib/rdflib), Apache Jena (https://jena.apache.org), and so on. The licenses of Protégé, RDFLib, and Apache Jena are

BSD 2-Clause, BSD 3-Clause and Apache License 2.0, respectively; using Stardog Studio requires for a license from the developers.

Code availability

All custom codes used in this study have been uploaded to Github[93].

Received: 15 August 2021; Accepted: 18 March 2022;
Published online: 17 June 2022

References

1. Permyakov, E. *Metalloproteomics*, 2 (John Wiley & Sons, 2009).
2. Bertini, G. *et al*. *Biological inorganic chemistry: structure and reactivity* (University Science Books, 2007).
3. Tamames, B., Sousa, S. F., Tamames, J., Fernandes, P. A. & Ramos, M. J. Analysis of zinc-ligand bond lengths in metalloproteins: trends and patterns. *Proteins: Structure, Function, and Bioinformatics* **69**, 466–475 (2007).
4. Sala, D., Giachetti, A. & Rosato, A. Molecular dynamics simulations of metalloproteins: A folding study of rubredoxin from Pyrococcus furiosus. *AIMS Biophys* **5**, 77–96 (2018).
5. Zhou, M. *et al*. A novel calcium-binding site of von Willebrand factor A2 domain regulates its cleavage by ADAMTS13. *Blood* **117**, 4623–4631 (2011).
6. Gogoi, P., Chandravanshi, M., Mandal, S. K., Srivastava, A. & Kanaujia, S. P. Heterogeneous behavior of metalloproteins toward metal ion binding and selectivity: insights from molecular dynamics studies. *Journal of Biomolecular Structure and Dynamics* **34**, 1470–1485 (2016).
7. Baldauf, C. *et al*. How cations change peptide structure. *Chemistry–A European Journal* **19**, 11224–11234 (2013).
8. De, S., Musil, F., Ingram, T., Baldauf, C. & Ceriotti, M. Mapping and classifying molecules from a high-throughput structural database. *Journal of Cheminformatics* **9**, 1–14 (2017).
9. Ropo, M., Blum, V. & Baldauf, C. Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead. *Scientific Reports* **6**, 1–11 (2016).
10. Vitalini, F., Mey, A. S., Noé, F. & Keller, B. G. Dynamic properties of force fields. *The Journal of Chemical Physics* **142**, 02B611_1 (2015).
11. Schneider, M. & Baldauf, C. Relative energetics of acetyl-histidine protomers with and without Zn^{2+} and a benchmark of energy methods. *arXiv preprint arXiv:1810.10596* (2018).
12. Maksimov, D., Baldauf, C. & Rossi, M. The conformational space of a flexible amino acid at metallic surfaces. *International Journal of Quantum Chemistry* **121**, e26369 (2021).
13. Marianski, M., Supady, A., Ingram, T., Schneider, M. & Baldauf, C. Assessing the accuracy of across-the-scale methods for predicting carbohydrate conformational energies for the examples of glucose and α-maltose. *Journal of Chemical Theory and Computation* **12**, 6157–6168 (2016).
14. Wang, J. & Kollman, P. A. Automatic parameterization of force field by systematic search and genetic algorithms. *Journal of Computational Chemistry* **22**, 1219–1228 (2001).
15. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry* **25**, 1656–1676 (2004).
16. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **118**, 11225–11236 (1996).
17. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **21**, 1049–1074 (2000).
18. Riniker, S. Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview. *Journal of Chemical Information and Modeling* **58**, 565–578 (2018).
19. Shivakumar, D., Harder, E., Damm, W., Friesner, R. A. & Sherman, W. Improving the prediction of absolute solvation free energies using the next generation opls force field. *Journal of chemical theory and computation* **8**, 2553–2558 (2012).
20. Allen, T. W., Andersen, O. S. & Roux, B. Energetics of ion conduction through the gramicidin channel. *Proceedings of the National Academy of Sciences* **101**, 117–122 (2004).
21. Roca, M. *et al*. Theoretical modeling of enzyme catalytic power: analysis of "cratic" and electrostatic factors in catechol O-methyltransferase. *Journal of the American Chemical Society* **125**, 7726–7737 (2003).
22. Zeng, J., Jia, X., Zhang, J. Z. & Mei, Y. The F130L mutation in streptavidin reduces its binding affinity to biotin through electronic polarization effect. *Journal of Computational Chemistry* **34**, 2677–2686 (2013).
23. Li, Y. L., Mei, Y., Zhang, D. W., Xie, D. Q. & Zhang, J. Z. Structure and dynamics of a dizinc metalloprotein: effect of charge transfer and polarization. *The Journal of Physical Chemistry B* **115**, 10154–10162 (2011).
24. Xie, W., Pu, J. & Gao, J. A coupled polarization-matrix inversion and iteration approach for accelerating the dipole convergence in a polarizable potential function. *The Journal of Physical Chemistry A* **113**, 2109–2116 (2009).
25. Ngo, V. *et al*. Quantum effects in cation interactions with first and second coordination shell ligands in metalloproteins. *Journal of Chemical Theory and Computation* **11**, 4992–5001 (2015).
26. Amin, K. S. *et al*. Benchmarking polarizable and non-polarizable force fields for Ca^{2+}–peptides against a comprehensive QM dataset. *The Journal of Chemical Physics* **153**, 144102 (2020).
27. Liang, G., Fox, P. C. & Bowen, J. P. Parameter analysis and refinement toolkit system and its application in MM3 parameterization for phosphine and its derivatives. *Journal of Computational Chemistry* **17**, 940–953 (1996).
28. Faller, R., Schmitz, H., Biermann, O. & Müller-Plathe, F. Automatic parameterization of force fields for liquids by simplex optimization. *Journal of Computational Chemistry* **20**, 1009–1017 (1999).
29. Cisneros, G. A., Karttunen, M., Ren, P. & Sagui, C. Classical electrostatics for biomolecular simulations. *Chemical Reviews* **114**, 779–814 (2014).
30. Rezac, J., Bm, D., Gutten, O. & Rulisek, L. Toward accurate conformational energies of smaller peptides and medium-sized macrocycles: MPCONF196 benchmark energy data set. *Journal of Chemical Theory and Computation* **14**, 1254–1266 (2018).
31. Jurečka, P., Šponer, J., Černý, J. & Hobza, P. 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. *Physical Chemistry Chemical Physics* **8**, 1985–1993 (2006).
32. Goerigk, L. *et al*. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Physical Chemistry Chemical Physics* **19**, 32184–32215 (2017).
33. Dohm, S., Hansen, A., Steinmetz, M., Grimme, S. & Checinski, M. P. Comprehensive thermochemical benchmark set of realistic closed-shell metal organic reactions. *Journal of Chemical Theory and Computation* **14**, 2596–2608 (2018).
34. Yu, W. *et al*. Extensive conformational searches of 13 representative dipeptides and an efficient method for dipeptide structure determinations based on amino acid conformers. *Journal of Computational Chemistry* **30**, 2105–2121 (2009).

35. Kishor, S., Dhayal, S., Mathur, M. & Ramaniah, L. M. Structural and energetic properties of α-amino acids: A first principles density functional study. *Molecular Physics* **106**, 2289–2300 (2008).
36. Selvarengan, P. & Kolandaivel, P. Potential energy surface study on glycine, alanine and their zwitterionic forms. *Journal of Molecular Structure: THEOCHEM* **671**, 77–86 (2004).
37. Császár, A. G. & Perczel, A. Ab initio characterization of building units in peptides and proteins. *Progress in Biophysics and Molecular Biology* **71**, 243–309 (1999).
38. Schlund, S., Müller, R., Grassmann, C. & Engels, B. Conformational analysis of arginine in gas phase–A strategy for scanning the potential energy surface effectively. *Journal of Computational Chemistry* **29**, 407–415 (2008).
39. Riffet, V., Frison, G. & Bouchoux, G. Acid–base thermochemistry of gaseous oxygen and sulfur substituted amino acids (Ser, Thr, Cys, Met). *Physical Chemistry Chemical Physics* **13**, 18561–18580 (2011).
40. Baek, K., Fujimura, Y., Hayashi, M., Lin, S. & Kim, S. Density functional theory study of conformation-dependent properties of neutral and radical cationic L-tyrosine and L-tryptophan. *The Journal of Physical Chemistry A* **115**, 9658–9668 (2011).
41. Floris, F. M., Filippi, C. & Amovilli, C. A density functional and quantum Monte Carlo study of glutamic acid in vacuo and in a dielectric continuum medium. *The Journal of Chemical Physics* **137**, 075102 (2012).
42. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **8**, 3192–3203 (2017).
43. Ropo, M., Schneider, M., Baldauf, C. & Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Scientific Data* **3**, 1–13 (2016).
44. Huang, H., Li, D. & Cowan, J. Biostructural chemistry of magnesium. regulation of mithramycin-DNA interactions by Mg^{2+} coordination. *Biochimie* **77**, 729–738 (1995).
45. Romani, A. M. Cellular magnesium homeostasis. *Archives of biochemistry and biophysics* **512**, 1–23 (2011).
46. Forsen, S. & Kordel, J. Calcium in biological systems (1994).
47. Grauffel, C., Dudev, T. & Lim, C. Why cellular di/triphosphates preferably bind Mg^{2+} and not Ca^{2+}. *Journal of Chemical Theory and Computation* **15**, 6992–7003 (2019).
48. Mahmoud, W. E. Functionalized ME-capped CdSe quantum dots based luminescence probe for detection of Ba^{2+} ions. *Sensors and Actuators B: Chemical* **164**, 76–81 (2012).
49. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 1–9 (2016).
50. Wittenburg, P., Lautenschlager, M., Thiemann, H., Baldauf, C. & Trilsbeek, P. FAIR practices in Europe. *Data Intelligence* **2**, 257–263 (2020).
51. Noy, N. F., *et al*. Ontology development 101: A guide to creating your first ontology (2001).
52. Wales, D. J. & Doye, J. P. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A* **101**, 5111–5116 (1997).
53. Wales, D. J. & Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **285**, 1368–1372 (1999).
54. Blum, V. *et al*. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **180**, 2175–2196 (2009).
55. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient O (N) integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics* **228**, 8367–8379 (2009).
56. Ren, X. *et al*. Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New Journal of Physics* **14**, 053020 (2012).
57. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865 (1996).
58. Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Physical Review Letters* **102**, 073005 (2009).
59. Swendsen, R. H. & Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters* **57**, 2607 (1986).
60. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**, 141–151 (1999).
61. Wong, M. A. & Hartigan, J. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**, 100–108 (1979).
62. Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoretica Chimica Acta* **44**, 129–138 (1977).
63. DiStasio, R. A., Gobre, V. V. & Tkatchenko, A. Many-body van der Waals interactions in molecules and condensed matter. *Journal of Physics: Condensed Matter* **26**, 213202 (2014).
64. Henkelman, G., Arnaldsson, A. & Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Computational Materials Science* **36**, 354–360 (2006).
65. Sanville, E., Kenny, S. D., Smith, R. & Henkelman, G. Improved grid-based algorithm for Bader charge allocation. *Journal of Computational Chemistry* **28**, 899–908 (2007).
66. Yu, M. & Trinkle, D. R. Accurate and efficient algorithm for Bader charge integration. *The Journal of Chemical Physics* **134**, 064111 (2011).
67. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993).
68. Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **5**, 129–145 (1984).
69. Fox, T. & Kollman, P. A. Application of the RESP methodology in the parametrization of organic solvents. *The Journal of Physical Chemistry B* **102**, 8070–8079 (1998).
70. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Antechamber: an accessory software package for molecular mechanical calculations. *J. Am. Chem. Soc* **222**, U403 (2001).
71. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**, 198–210 (2013).
72. O'Boyle, N. M. *et al*. Open Babel: An open Chemical toolbox. *Journal of Cheminformatics* **3**, 1–14 (2011).
73. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**, 1859–1865 (2008).
74. Eastman, P. *et al*. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology* **13**, e1005659 (2017).
75. Hu, X. & Baldauf, C. Cation-coordinated conformers of 20 proteinogenic amino acids with different protonation states. *NOMAD* https://doi.org/10.17172/NOMAD/2021.02.10-1 (2021).
76. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials* **2**, 036001 (2019).
77. Hu, X., Lenz-Himmer, M. O. & Baldauf, C. The ontology representation for a data set of cation-coordinated conformers of 20 proteinogenic amino acids with different protonation states. *EDMOND* https://doi.org/10.17617/3.5q (2021).
78. Al-Aswadi, F. N., Chan, H. Y. & Gan, K. H. Automatic ontology construction from text: a review from shallow to deep learning trend. *Artificial Intelligence Review* **53**, 3901–3928 (2020).
79. Musen, M. A. The protégé project: a look back and a look forward. *AI Matters* **1**, 4–12 (2015).

80. Lamy, J.-B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine* **80**, 11–28 (2017).
81. Tsarkov, D. & Horrocks, I. FaCT++ description logic reasoner: System description. In *International Joint Conference on Automated Reasoning*, 292–297 (Springer, 2006).
82. Wang, J. *et al*. Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *The Journal of Physical Chemistry B* **116**, 7088–7101 (2012).
83. Li, Y. *et al*. Machine learning force field parameters from ab initio data. *Journal of Chemical Theory and Computation* **13**, 4492–4503 (2017).
84. Cole, D. J., Vilseck, J. Z., Tirado-Rives, J., Payne, M. C. & Jorgensen, W. L. Biomolecular force field parameterization via atoms-in-molecule electron density partitioning. *Journal of Chemical Theory and Computation* **12**, 2312–2323 (2016).
85. Rai, B. K. & Bakken, G. A. Fast and accurate generation of ab initio quality atomic charges using nonparametric statistical regression. *Journal of Computational Chemistry* **34**, 1661–1671 (2013).
86. Bleiziffer, P., Schaller, K. & Riniker, S. Machine learning of partial charges derived from high-quality quantum-mechanical calculations. *Journal of Chemical Information and Modeling* **58**, 579–590 (2018).
87. Møller, C. & Plesset, M. S. Note on an approximation treatment for many-electron systems. *Physical Review* **46**, 618 (1934).
88. Head-Gordon, M., Pople, J. A. & Frisch, M. J. MP2 energy evaluation by direct methods. *Chemical Physics Letters* **153**, 503–506 (1988).
89. Ambrosetti, A., Reilly, A. M., DiStasio, R. A. Jr & Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *The Journal of Chemical Physics* **140**, 18A508 (2014).
90. Riplinger, C. & Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *The Journal of Chemical Physics* **138**, 034106 (2013).
91. Riplinger, C., Sandhoefer, B., Hansen, A. & Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *The Journal of Chemical Physics* **139**, 134101 (2013).
92. Supady, A., Blum, V. & Baldauf, C. First-principles molecular structure search with a genetic algorithm. *Journal of Chemical Information and Modeling* **55**, 2338–2348 (2015).
93. Hu, X. XiaojuanHu/AA_property_calculation: First release of AA_property_calculation. *Zenodo* https://doi.org/10.5281/zenodo.5672781 (2021).

Acknowledgements

X.H. is grateful for a doctoral fellowship by the China Scholarship Council. All authors acknowledge funding by the Federal Ministry of Education and Research of Germany for the project STREAM ("Semantische Repräsentation, Vernetzung und Kuratierung von qualitätsgesicherten Materialdaten", ID: 16QK11C).

Author contributions

X.H. performed the calculations of all conformers, curated the data, constructed the ontology, and wrote the manuscript. M.L. helped with the construction of ontology and contributed to the manuscript. C.B. designed the study, curated the data, and wrote the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.H. or C.B.

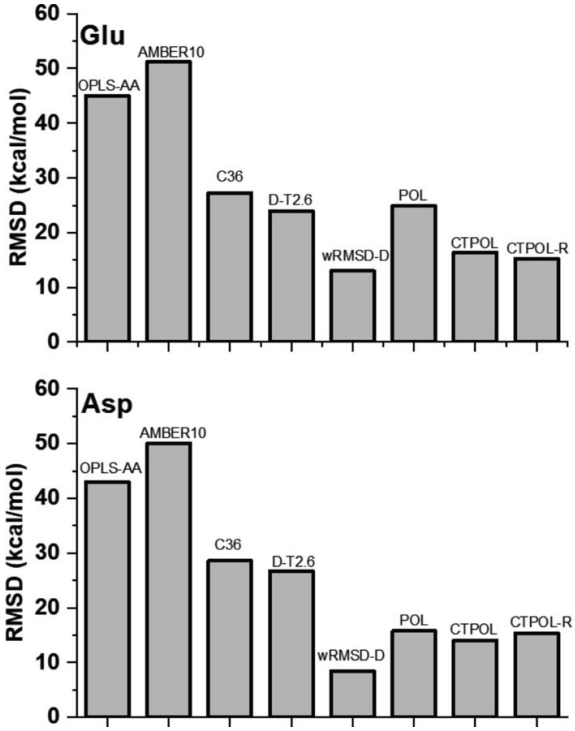
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.2 Paper II: Benchmarking polarizable and non-polarizable force fields for Ca^{2+}–peptides against a comprehensive QM dataset

K. S. Amin, **X. Hu**, D. R. Salahub, C. Baldauf, C. Lim and S. Noskov.

J. Chem. Phys. ***153****, 144102 (2020)*



DOI: 10.1063/5.0020768

Author contributions: K.S. Amin and I contributed equally to this work. C. Baldauf and I calculated and organized the QM data set. I assessed the accuracy of OPLS-AA and Amber by their abilities to reproduce hierarchies of thousands of Ca^{2+}–dipeptide interaction energies in the QM data set. K.S. Amin assessed the accuracy of CHARMM36 and Drude model. K.S. Amin optimized the selected cation–peptide parameters in Drude model. I parameterized and evaluated the CTPOL model. K.S. Amin, D.R. Salahub and S. Noskov did the MD simulations in condensed-phase.

Benchmarking polarizable and non-polarizable force fields for Ca^{2+}–peptides against a comprehensive QM dataset

Cite as: J. Chem. Phys. **153**, 144102 (2020); https://doi.org/10.1063/5.0020768
Submitted: 06 July 2020 • Accepted: 18 September 2020 • Published Online: 08 October 2020

Kazi S. Amin, Xiaojuan Hu, Dennis R. Salahub, et al.

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Benchmarking polarizable and non-polarizable force fields for Ca^{2+}–peptides against a comprehensive QM dataset

Cite as: J. Chem. Phys. 153, 144102 (2020); doi: 10.1063/5.0020768
Submitted: 6 July 2020 • Accepted: 18 September 2020 •
Published Online: 8 October 2020

Kazi S. Amin,[1] Xiaojuan Hu,[2] Dennis R. Salahub,[3] Carsten Baldauf,[2] Carmay Lim,[4,5,a)] and Sergei Noskov[1,a)]

AFFILIATIONS

[1]CMS – Centre for Molecular Simulation and Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada
[2]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany
[3]Department of Chemistry, CMS – Centre for Molecular Simulation, IQST – Institute for Quantum Science and Technology, Quantum Alberta, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada
[4]Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan
[5]Department of Chemistry, National Tsing Hua University, Hsinchu 300, Taiwan

Note: This paper is part of the JCP Special Topic on Classical Molecular Dynamics (MD) Simulations: Codes, Algorithms, Force Fields, and Applications.
[a)]**Authors to whom correspondence should be addressed:** carmay@gate.sinica.edu.tw and snoskov@ucalgary.ca

ABSTRACT

Explicit description of atomic polarizability is critical for the accurate treatment of inter-molecular interactions by force fields (FFs) in molecular dynamics (MD) simulations aiming to investigate complex electrostatic environments such as metal-binding sites of metalloproteins. Several models exist to describe key monovalent and divalent cations interacting with proteins. Many of these models have been developed from ion–amino-acid interactions and/or aqueous-phase data on cation solvation. The transferability of these models to cation–protein interactions remains uncertain. Herein, we assess the accuracy of existing FFs by their abilities to reproduce hierarchies of thousands of Ca^{2+}–dipeptide interaction energies based on density-functional theory calculations. We find that the Drude polarizable FF, prior to any parameterization, better approximates the QM interaction energies than any of the non-polarizable FFs. Nevertheless, it required improvement in order to address polarization catastrophes where, at short Ca^{2+}–carboxylate distances, the Drude particle of oxygen overlaps with the divalent cation. To ameliorate this, we identified those conformational properties that produced the poorest prediction of interaction energies to reduce the parameter space for optimization. We then optimized the selected cation–peptide parameters using Boltzmann-weighted fitting and evaluated the resulting parameters in MD simulations of the N-lobe of calmodulin. We also parameterized and evaluated the CTPOL FF, which incorporates charge-transfer and polarization effects in additive FFs. This work shows how QM-driven parameter development, followed by testing in condensed-phase simulations, may yield FFs that can accurately capture the structure and dynamics of ion–protein interactions.

I. INTRODUCTION

Molecular dynamics (MD) simulations are making great strides in research on biomolecular phenomena. This is largely due to increased computational power and superior numerical techniques, which allow researchers to model and simulate a variety of large biomolecular systems on experimentally accessible time scales of milli-seconds.[1–4] We can now exploit higher computational efficiency to incorporate much needed theoretical improvements, broadening the applicability of MD models for the next generation

of biomolecular research.[3,5,6] The majority of current MD simulation studies rely on classical force fields (FFs) such as CHARMM,[4] AMBER,[7] GROMOS,[8] and OPLS-AA.[9] However, these additive FF models fail to provide sufficient accuracy for several important biological systems, particularly those involving crucial metal–protein interactions.[3,6,10–16] One of the major limitations in the otherwise successful additive FF approximation is the lack of explicit treatment of an atom's electronic degrees of freedom, a crucial determinant of realistic molecular behavior in metalloprotein systems, especially those with divalent cations. Although additive FF refinements such as ECCR,[17–19] adaptive force-matching algorithms utilizing *ab initio* energies for the refinement of additive force fields,[20,21] or the 12-6-4 form of the Lennard-Jones (LJ) potential have been successful to a degree in this regard,[16,22,23] they are still limited in their scope due to the diversity of electrostatic environments found in proteins.

An alternative approach is to account for the polarization of each atom explicitly in the general molecular mechanics (MM) framework.[3,6,24–26] There is strong and rapidly growing evidence that in many cases, polarizable FFs reproduce experimental thermodynamics data as well as high-level quantum mechanical (QM) results more accurately than fixed-charge models. For instance, compared with fixed-charge models, they predict better ion solvation enthalpies and free energies,[3,27–30] protein–ligand recognition and binding,[3,6] and the pK_a of amino-acid residues in water and protein environments.[31] The explored approaches vary from the implementation of fluctuating charge schemes to models relying on the induced-dipole approximation, each with apparent advantages but also with caveats. Fluctuating charge (FQ) models simulate charge transfer dynamically by redistributing the atomic charges to equalize electronegativity, while keeping the total charge conserved.[25,26] Notable FQ models are CHARMM-FQ and ABEEMsp (atom-bond electronegativity equalization model with s- and p-bonds).[32,33] One of the major drawbacks of FQ models is that they fail to capture out-of-plane polarization effects, which are critical for describing many common functional groups such as aromatic rings. Attempts to include out-of-plane effects using virtual charge sites can also prove to be inefficient due to challenges in scaling to simulation systems containing thousands of atoms.[33]

Induced-dipole models explicitly account for polarizability by implementing a dynamic electric dipole that responds to changes in the surrounding electrostatic environment. Notable FFs that use this approximation are the CHARMM Drude oscillator model,[3] AMOEBA (atomic multipole optimized energetics for biomolecular simulation),[28,29] and SIBFA (sum of interactions between fragments *ab initio*) FFs.[24,34] Some of these methods can be expanded beyond dipolar approximations by including higher order multipole terms and also by accounting for charge transfer.[6,26]

One area that remains as a frontier for the development of polarizable FFs is the chemically accurate description of cation–protein interactions, particularly divalent ions such as Ca^{2+} and Mg^{2+}. Efforts in the last decade show that polarizable FFs model divalent ion–protein interactions more accurately than their non-polarizable counterparts. For instance, the AMOEBA polarizable FF has recently been used to predict more accurate relative binding free-energies and Ca^{2+} or Mg^{2+} selectivity of model soluble protein systems, where non-polarizable FFs fail even after extensive parameterization efforts.[35] Roux and colleagues[36] performed an exhaustive optimization of Drude parameters and showed the superior performance of Drude polarizable FFs in studies of aqueous salt solutions of monovalent and divalent cations. Li *et al.*[10] investigated the parameter space required to accurately describe gas-phase interaction energies between physiological cations and a set of protein binding sites. The gas-phase QM energies were used as a reference dataset to guide Drude FF development with ion–carboxylate interactions noted as a potential focus of parameter optimization. While the parameters were shown to provide excellent performance in various reduced models of binding sites,[11,37] their extension to MD simulations of ion–protein interactions and transport in porin proteins elucidated remarkable issues leading to a hindered ion diffusion in the protein interior as well as apparent over-binding to the protein.[38,39]

Recently, Villa *et al.* showed, using the Drude FF, that it is possible to capture the complex interaction surface of Mg^{2+} with methyl phosphate in the condensed phase, illustrating the feasibility of developing accurate and transferable polarizable potential functions for metal–ligand interactions.[40] However, success in the final deployment of next-generation polarizable FFs depends critically on assessing (i) the vast chemical space presented by the variety of side chains found in proteins and (ii) the strategies for explicitly including charge transfer terms in the case of strongly interacting cations. In metalloproteins containing strong charge donors such as negatively charged carboxylate or thiolate groups lining the metal-binding site, ligand → cation charge transfer is significant.[15] However, ligand → cation charge transfer reduces the magnitude of partial charges on the metal-ligating atoms and cation in conventional additive FFs, thus attenuating their charge/dipole–charge interactions. This can be compensated by including the local polarization energies of the cation and its ligands. Based on these physical principles, Sakharov and Lim[41,42] developed the CTPOL FF, which incorporates charge transfer and local polarization effects directly into the additive potential functions, for metalloprotein simulations.

In this paper, we used data from a large set of structures and energies based on density-functional theory (DFT) calculations that was created by an exhaustive structure search by Ropo *et al.*[43,44] The dataset comprises the proteinogenic amino acids in various protonation states and their amino-methylated and acetylated (capped) dipeptides bound to Ca^{2+} and other divalent cations. Using such data, we follow a complementary alternative to the established molecular-fragment approach. Based on the ion–dipeptide geometries and interaction energies in the dataset, we compare the performance of the polarizable Drude FF with three fixed-charge FFs, namely, CHARMM (C36), AMBER, and OPLS-AA. The goal of this comparative study is to (i) assess the ability of modern FFs to accurately describe peptide–divalent cation interactions in a complex chemical space, (ii) reduce the chemical space for future FF development by locating atom-types of interest to provide insight into how we may reduce the parameter space for optimization, and (iii) assess the impact of the explicit account of charge-transfer (CT) and local polarization effects between the protein host and the bound cation using the CTPOL approach. First, we identified the chemical space where the Drude FF fails. We then show how this can be amended by parameterization of a few selected parameters using two different objective functions. By relating parameter space to conformational space, we illustrate the utility of first-principles methods such as DFT as a reference and the choice of objective function for the future optimization of polarizable FFs.

II. THEORY AND METHODS

A. Cation–dipeptide reference structures

The DFT-based references were built from a large dataset of dipeptide structures, as depicted in Fig. 1, where **R** represents an amino-acid side chain. The dataset includes various cations and bare amino acids and dipeptides, with over 45 000 stationary points on the respective potential-energy surfaces.[43,44] In this paper, we studied only the Ca^{2+}-bound dipeptides with a total of 2583 conformations. The conformations and total energies of each molecular system were calculated using the Perdew–Burke–Ernzerhof (PBE) generalized-gradient exchange-correlation functional, chosen after testing several other functionals.[45,46] Energies were corrected for van der Waals interactions using the Tkatchenko–Scheffler formalism.[47] PBE with a pairwise dispersion correction represents a good compromise between accuracy and computational cost. This choice of the functional was validated in the original dataset article.[43,44,46] Furthermore, the generalized gradient approximation (GGA) functional PBE has been shown to produce acceptable mean-absolute errors in comparison to coupled cluster calculations for related systems.[46] The focus of the cited work is to check whether one can represent the complexity of an all-electron approach with an extended force field; thus, "any" DFT method would suffice.

All the electronic structure calculations were carried out using the numeric atom-centered basis set all-electron code FHI-aims.[48] The standard *tight* settings of FHI-aims for all species were used. The initial global conformational search was performed by a basin hopping search strategy using the OPLS-AA FF,[9] and the energy minima identified were subsequently relaxed using PBE+vdW with *light* settings.[49] The identified set of structures was then subjected to a further first-principles refinement step by *ab initio* molecular dynamics with replica-exchange to enhance sampling.[50] The obtained conformers were further relaxed using PBE+vdW (*tight* settings) and clustered using a *k*-means clustering algorithm with a cluster radius of 0.3 Å to obtain the final conformation hierarchies.[51] The dataset shows good agreement with available experimental data for gas-phase ion affinities.[43,44] A two-stage restrained electrostatic potential (RESP) fitting procedure was employed to obtain partial atomic charges for various ion–peptide conformations based on electrostatic potentials calculated with FHI-aims[48] at the level of theory described above. RESP calculations were performed on a radial grid of point charges fixed in a cubic space around the ion–peptide complex. The 5 radial shells of point charges were generated in a region between 1.4 and 2.0 multiples of the atomic vdW radius. The cubic grid for RESP calculations contained 35 point charges along x, y, and z directions, respectively, to assess the electrostatic potential (ESP) around the ion–peptide complex. The Antechamber suite of the AmberTools package[7] was used for RESP charge fitting.[52]

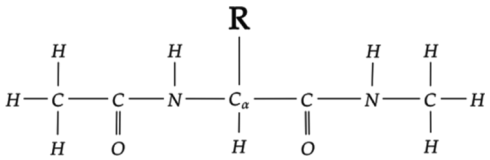


FIG. 1. Structure of a dipeptide, with a variable side chain (**R**) that extends from the alpha-carbon (C_α), which can be any one of the 20 proteinogenic side chains, plus a few variations of His, namely, HSD (with hydrogen on N^δ), HSE (with hydrogen on N^ε), and HSP (with hydrogens on both nitrogens), which are the standard protonation states found in C36 and Drude FFs.

B. Additive force fields for peptide–ion interactions

All the additive FFs used in this study rely on the fixed-charge representation illustrated in Fig. 2. In additive FFs, atoms are represented as hard spheres with point charges ("balls" in the figure) and bonds as springs ("sticks" in the figure) with a number of intra-molecular terms to account for bond, angular and dihedral-improper degrees of freedom.

The intra- and inter-molecular interactions in a polyatomic system can be described by a potential-energy function given by

$$U_{FF} = \sum_{bonds} K_b(b-b_0)^2 + \sum_{1-3bonds} K_{UB}(S-S_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedrals,n} K_{\psi,n}[1+\cos(n\psi-\delta_n)] + \sum_{improper} K_\chi(\chi-\chi_0)^2 + \sum_{i<j}\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}\frac{q_i q_j}{4\pi\epsilon r_{ij}}. \quad (1)$$

In Eq. (1), K, b_0, θ_0, S_0, χ_0, n, δ_n, ε_{ij}, σ_{ij}, and q are empirically determined parameters. The force constants [K and parameters of the harmonic terms (b_0, S_0, θ_0, χ_o, n, and δ_n)] are usually obtained using analysis of QM vibrational modes. The partial charges q_i are generally obtained by fitting to electrostatic potential surfaces obtained via QM. After determining the bonded parameters and partial charges, the Lennard-Jones (LJ) terms (ε_{ij}, σ_{ij}) are finally fitted to reproduce both gas-phase QM energies and condensed-phase thermodynamics such as experimental hydration free energies.

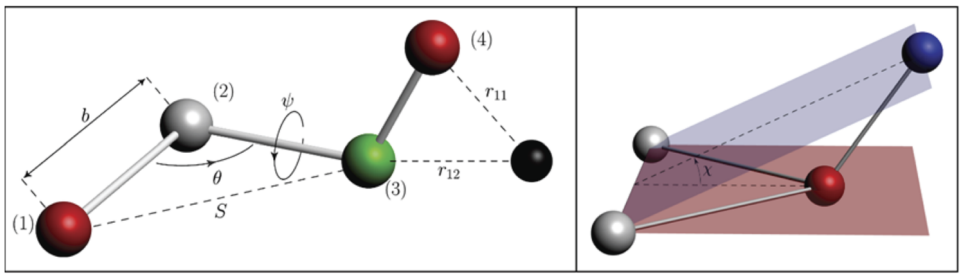


FIG. 2. Ball and stick model of classical FFs. Left: configuration of a proper dihedral. Right: configuration of an improper dihedral.

In the present study, we examine the accuracies of the following popular additive FFs: OPLS-AA,[9] AMBER,[7] CHARMM36m,[4] and CHARMM Drude FF with latest protein parameters[3,53] (see Sec. II C). OPLS-AA and AMBER employ a functional form similar to that used by CHARMM, except that (i) OPLS-AA and AMBER do not use the Urey–Bradley (UB) form for the intra-molecular angular potential, and (ii) in AMBER and OPLS-AA, the standard dihedral-angle torsion term is used for the out-of-plane distortions, which corresponds to the improper term in Eq. (1). The AMBER FF used in this work is AMBER10.[54] OPLS-AA and AMBER data in this paper were calculated using openMM7, a high performance toolkit for molecular simulations.[55] The CHARMM36m FF[4] used to model dipeptide–cation interactions was used with a set of non-bonded fix (NBFIX) terms directly from the CHARMM-GUI portal without any modifications.[56]

C. The Drude polarizable force field

In the Drude polarizable FF, an additional particle is attached to every polarizable (heavy) atom, as depicted in Fig. 3. This particle is assigned to a point partial charge and a constant mass of 0.4 amu or 0.8 amu. The spring constant may also be a non-diagonal tensor, which can capture anisotropic polarizability. The lone-pair particles are used to better represent the charge distribution in diverse chemical groups found in proteins. The auxiliary Drude particles are included in the extended Lagrangian framework[57] and added to the set of particles that contribute to the Coulomb electrostatic energy in Eq. (1). They also contribute energy due to displacement from their host nuclei, given by Eq. (2),

$$E_D = \frac{1}{2}\sum_p K_{D,p} d_p^2. \tag{2}$$

D. Electrostatic interactions and polarization catastrophe

The transfer of the developed parameters for metalloproteins to condensed-phase simulations is complicated by several issues including polarization catastrophes as well as the limited set of protein sites used by Li *et al.*[10] The polarization catastrophe or over-polarization phenomenon is due to fundamental differences between QM and polarizable FFs, which neglect electron–electron overlap and charge-transfer effects. When a polarizable atom is close to a charged atom or another highly polarizable atom, one or both of them may over-polarize, and the mutual inductance of dipoles can cause a chain reaction that induces over-polarization of other atoms, thus amplifying the effect. The phenomenon has been observed in systems with high charge density[4] and has been documented previously with the Drude polarizable FF, especially when divalent ions and charged moieties are involved.[3] A popular method for handling over-polarization in the Drude polarizable model relies on the implementation of a Thole damping function that screens the Coulomb potential at short distances.[58] The Thole function, Eq. (3), effectively screens the electrostatic interaction at short distances, leaving the long-range interactions untouched, using a distance-dependent function,[59]

$$S_{ij}(r) = 1 - \left(1 + \frac{t_{ij} r}{2(\alpha_i \alpha_j)^{\frac{1}{6}}}\right)\exp\left[\frac{-t_{ij} r}{(\alpha_i \alpha_j)^{\frac{1}{6}}}\right]. \tag{3}$$

In Eq. (3), t_{ij} is a pair-specific Thole factor between atoms i and j, α are the atomic polarizabilities, and r is the interatomic distance. The damping effect applies not only to the atomic nuclei but also to the Drude particles. This prevents a polarization catastrophe at short distances, while maintaining the electrostatic interactions at longer distances. The effective distance and strength of damping are controlled by the Thole factor, t_{ij}.

The effect of such a function on the Coulomb potential is depicted in Fig. 4. Essentially, it corrects the Coulomb potential to treat the atom as if it were a smeared charge distribution, removing the singularity of a point charge. Though Thole damping is effective against over-polarization, it contains some inadequacies because it does not account for many-body polarization effects.[60] However, optimizing the exponent of the Thole function (t_{ij}) may improve the accuracy[61] and aid the description of cation–peptide interaction energies.

E. Assessment of the cation–dipeptide interaction energies

The ion–dipeptide interaction energies for all additive models obtained on the basis of QM geometries from the dipeptide dataset use infinite cutoffs. The ion–dipeptide interaction energies for polarizable models were obtained by relaxing the Drude particles via

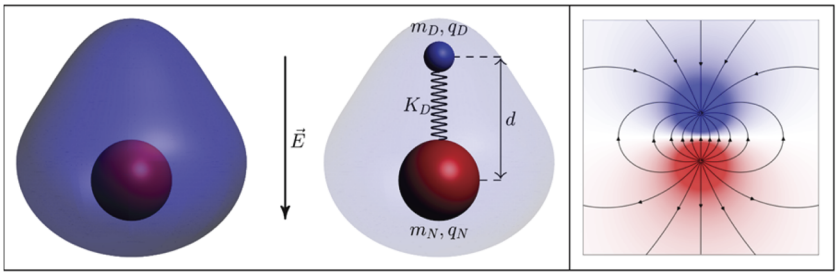


FIG. 3. Schematic of a Drude polarizable atom and resulting FF. Left: conceptual depiction of electron density around an atom polarized by an external electric field, $\vec{E}$. Middle: Drude model of the same atom, with Drude particles in blue. The Drude particles have a mass of m_D and charge of q_D, whereas the parent nucleus has a mass of m_N and a charge of q_N. The distance d is controlled by a spring with the force constant K_D. Right: a physical dipole with field lines representing potential gradients.

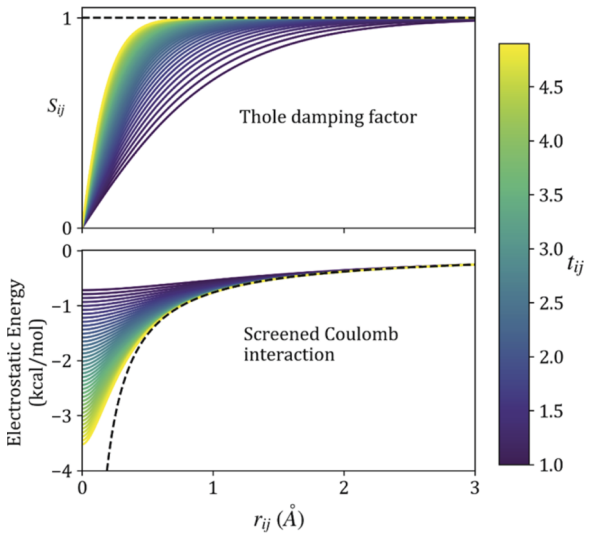


FIG. 4. Effect of Thole damping function for various t_{ij}. Each curve represents a particular screening factor (t_{ij}) where the color represents its value. The top panel shows S_{ij} as a function of distance r, and the bottom panel shows the corresponding screened Coulomb potential energy (kcal/mol). The dashed line represents the infinite Thole limit, where there is no screening [$S_{ij}(r) = 1\ \forall\ r$].

steepest-descent for 500 steps followed by adopted-basis Newton–Raphson minimization for 100 steps to a final gradient of 10^{-5} kcal mol^{-1} Å^{-1}, with the atomic positions restrained by a force constant of 10^{7} kcal mol^{-1} Å^{-2} to the reference QM geometry from the dipeptide dataset. To evaluate the accuracy of the different FFs, we calculated the root-mean-squared deviation (RMSD) for each ion-bound dipeptide as follows:

$$RMSD = \left[\frac{1}{N}\sum_{i}\left(E^{i}_{QM} - E^{i}_{MM}\right)^{2}\right]^{\frac{1}{2}}, \tag{4}$$

where N is the total number of conformations and E^{i} is the interaction energy of the ith conformation.

F. Charge transfer modeling with the CTPOL FF

The CTPOL model[41,42] incorporates charge transfer and local polarization effects into additive force fields by modifying the conventional Coulombic term to account for ligand → cation charge transfer and including an additional term in the potential function (see below) to account for the induced polarization due to the bound cation. It enables incorporation of partial-charge transfer and induced polarization effects into an existing additive potential function as follows:

$$U^{CTPOL}_{Nonbonded} = E_{vdW} + E^{CT}_{stat} + E_{pol}. \tag{5}$$

The electrostatic interactions in CTPOL include dynamic charge transfer between the bound cation and atoms comprising its coordination shell (O,S, and N). The amount of charge transferred by a metal-ligating atom (L) to a metal cation (Me) is assumed to depend linearly on the interatomic distance, r_{Me-L}, and is given by

$$\Delta q_{Me-L} = a_L r_{Me-L} + b_L. \tag{6}$$

The a_L and b_L coefficients in Eq. (6) were obtained using Particle Swarm Optimization (PSO) and reproducing the relative QM interaction energies as the objective function. PSO relies on a population of solutions, called particles, which move through the high-dimensional parameter space with directed velocity vectors to find optimal solutions.[62,63] PSO was performed via the python package pyswarm.[64] The amount of charge transferred, Δq_{Me-L}, is added to the partial charge on atom L from a given classical FF to yield the net partial charge on atom L at any given simulation time step, t,

$$q_L = q^{0}_{L} + \Delta q_{Me-L}. \tag{7}$$

The polarization energy E_{pol} can be computed according to

$$E^{pol} = -\frac{1}{2}\sum_{i}\mu_i \cdot \boldsymbol{E}^{0}_{i}, \tag{8}$$

where the summation is over the cation and the metal-ligating amino-acid heavy atoms, μ_i is the dipole induced on atom i, and $\boldsymbol{E}^{0}_{i}$ is the electrostatic field produced by the current charges in the system at a polarizable site i. Polarizabilities of each atom type are taken as the average value of all corresponding effective atomic polarizabilities from the DFT data. Following previous work,[41,42] we employ a cutoff distance r_{ij}^{cutoff} equal to the sum of the vdW radii of atoms i and j scaled by a parameter $\gamma = 0.92$ so that interatomic distances $r_{ij} \le r_{ij}^{cutoff}$ are set equal to r_{ij}^{cutoff} to avoid unphysically high induced dipoles at close distances to each other and to the permanent electric charges. The additive AMBER10 FF[54] was used to describe dipeptides and long-range interactions between Ca^{2+} and dipeptides. The atom-type definitions for CTPOL developed in our work are shown in Fig. SI 1. The implementation and calculations of the CTPOL model were performed with openMM7.[55]

G. MD simulation protocol

To evaluate the performance of the various Drude polarizable FF parameters used in this paper, we ran MD simulations on the truncated structure of the N-lobe of the human calmodulin (CaM) protein (PDB 1CLL),[65] containing Ca^{2+}-bound EF-hand loops I and II. We used the CHARMM-GUI platform[56] to build a truncated CaM with Ca^{2+} bound to two characterized sites solvated in a neutralizing 150 mM $CaCl_2$ aqueous solution. The original crystal structure (1CLL) was solved in the acidic solution (pH = 5.0), containing 50 mM $MgCl_2$, 5 mM $CaCl_2$, and 50 mM NaOAc.[65] We chose a higher than physiological concentration of $CaCl_2$ to test ion interactions with the highly charged protein surface. The cubic simulation box (63.9 × 63.9 × 63.9 Å^3) contained 1 protein molecule, 28 Ca^{2+}, 43 Cl^{-}, and 8133 TIP3P water molecules.[66] The solvated system was first minimized using a staged-protocol of CHARMM-GUI for 60 ns (10 ns for each stage) using NAMD2.14b1,[67] with positional constraints applied to heavy protein atoms. The system was then simulated for 250 ns in a constant-pressure ensemble (NPT) at T = 298.15 K without any positional constraints using a time step of 2 fs. The electrostatic interactions were treated using the Particle Mesh Ewald (PME) method[68] with a grid spacing of 1 Å and

sixth-order interpolation with a real space cutoff of 12 Å. The LJ interactions were smoothly switched off from 10 Å–12 Å. The atom-pair list was updated every 20 steps. The LJ and electrostatic interactions were computed every time step. The SHAKE algorithm (RATTLE)[69] was used to maintain the geometry of all bonds involving hydrogen. The polarizable simulation system was built using a pre-equilibrated box described above with the CHARMM-GUI/Drude-Prepper option.[70] The latest Drude FF for proteins[53] was used with different Thole parameters for Ca^{2+}–O(carboxylate) interactions (as described in Sec. III).

Langevin dynamics with a dual-thermostat scheme was used to propagate the atoms and auxiliary Drude particles with the extended Lagrangian formalism implemented in the NAMD package.[67,71] The thermostat acting on heavy (non-Drude) particles was set to T_{atom} = 298.15 K. The Langevin damping coefficient was set to 5.0 ps^{-1}. Production runs of 250 ns were performed with T_{Drude} = 0.5 K and a spring constant for the atom–Drude bond of 1000 (kcal/mol)/Å for the different parameter sets considered in our work. The first 50 ns were discarded for all analyses shown in the text. A damping constant of 20.0 ps^{-1} was applied to Drude particles.[57] A "hard-wall" constraint was used to prevent large displacements of Drude particles in the case of strong electrostatic interactions expected in the simulation of divalent cation–protein interactions.[72,73] The hard-wall constraint distance was set to 0.2 Å, and a time step of 0.8 fs was used in all MD simulations performed with Drude FFs.

III. RESULTS AND DISCUSSION

A. Force-field performance in modeling ion–dipeptide interactions

The values of RMSD in cation–dipeptide interaction energies relative to the QM dataset and prior to any parameter optimization are plotted in Fig. 5. It is evident that the Drude polarizable FF is more accurate than the non-polarizable FFs for almost all the studied dipeptide structures, with average RMSDs significantly lower than 100 kcal/mol. When Ca^{2+} is in close proximity to charged carboxylate moieties, the auxiliary Drude particle of the oxygen atom is pulled onto the cation by undamped electrostatic forces [Fig. 6(a)], which causes the magnitude of the electrostatic energy to escalate above computational threshold values of 10^8 kcal/mol, resulting in unusually large RMSDs. However, even though the overall RMSD is better for Drude compared with C36, the lowest three conformations are in fact better captured by C36 [Fig. 6(b)]. The clear outliers for the Drude FF are the interactions of Ca^{2+} with negatively charged Asp and Glu side chains. Analysis of the outliers indicates that this discrepancy is caused by the over-polarization catastrophe phenomenon (see Sec. II D). Interestingly, although the polarization catastrophe in both Glu and Asp is due to the Drude–cation overlap, it occurs more frequently in Glu-dipeptide than in Asp-dipeptide, where it only occurs in one conformation. This is probably because the longer side chain of Glu allows a greater number of stable conformations, in which Ca^{2+} is close to the backbone oxygens and the carboxylate group. This appears to be a preferred coordination state for the ion when interacting with these dipeptides.

Figure 7 provides further details on the chemical space where the polarization catastrophe occurs. The Squared Difference (SD)

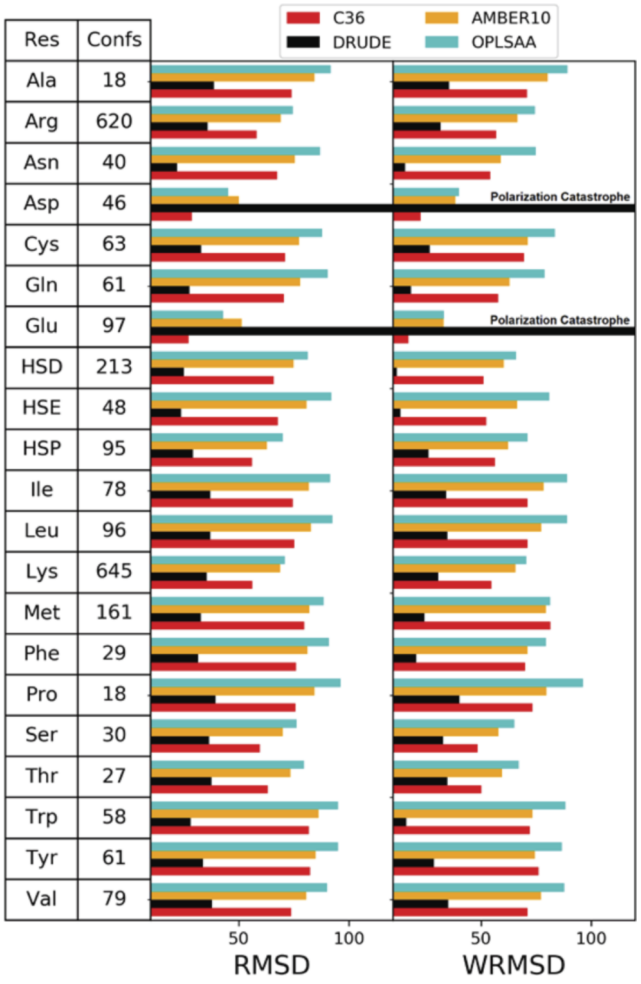


FIG. 5. Number of conformations and RMSD relative to the QM-interaction energies for each dipeptide residue. RMSDs are plotted for each of the four FFs prior to any optimization or correction. For Asp and Glu, the Drude RMSDs are on the order of 10^7 due to polarization catastrophe. wRMSD is the Boltzmann-weighted RMSD defined in Eqs. (9) and (10) (see below).

of the interaction energy is on the order of 10^{16} kcal/mol due to the polarization catastrophe in regions close to the two carboxylate oxygens. In Glu-dipeptide, the Drude FF evidently fails in the region where there is a significant electronic overlap (<2.2 Å) due to the polarization catastrophe discussed above. Although the average distance from ligating oxygen atoms to Ca^{2+} ranges from 2.37 Å to 2.41 Å,[74] a survey of high-resolution (<2.0 Å) PDB structures containing nonredundant Ca^{2+} sites reveals several structures with Ca–O distances <2.2 Å. Li *et al.*[10] used chemical structures of Ca^{2+}-containing peptides with average coordination distances of 2.39 Å in determining non-bonded parameters for Ca–O interactions and, therefore, have not considered conformations with a significant electronic overlap in their parameter determination. In particular, Figs. 6 and 7 highlight the significance of possible electron overlap between Ca^{2+} ions and the OE1, OE2, OD1, and OD2 atom

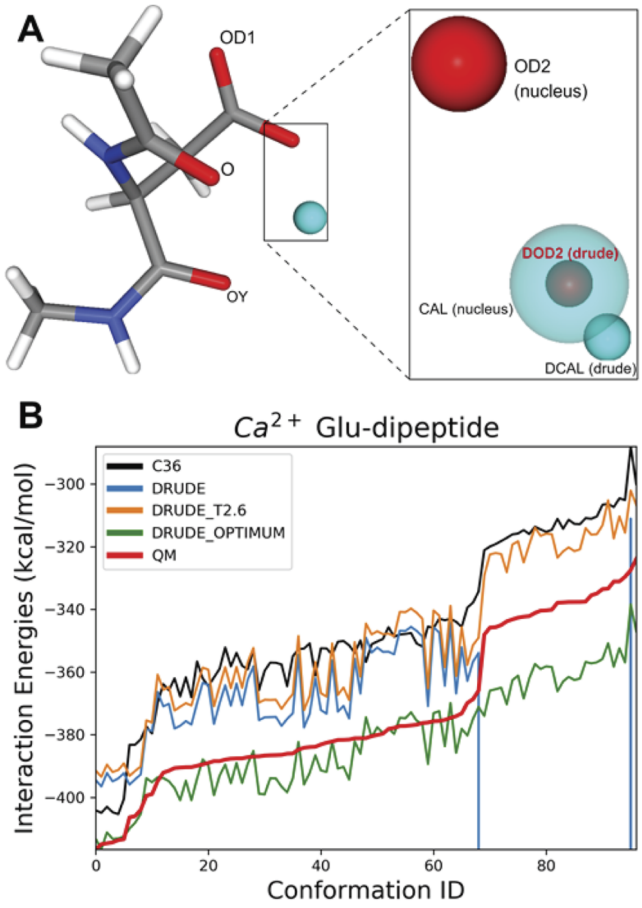


FIG. 6. (a) Polarization catastrophe in ASP dipeptide when Ca^{2+} is in close proximity to the OD2 atom. The Drude particle of OD2 (atom-type DOD2) is abnormally pulled away from its parent nucleus by the electrostatic force of Ca^{2+}, which also has its Drude particle (DCAL) abnormally far from its host nucleus. (b) Comparative analysis of conformation-specific dipeptide: Ca^{2+} interaction energies between various FFs and QM. For Drude, conformations 69–94 and 96 experience polarization catastrophe.

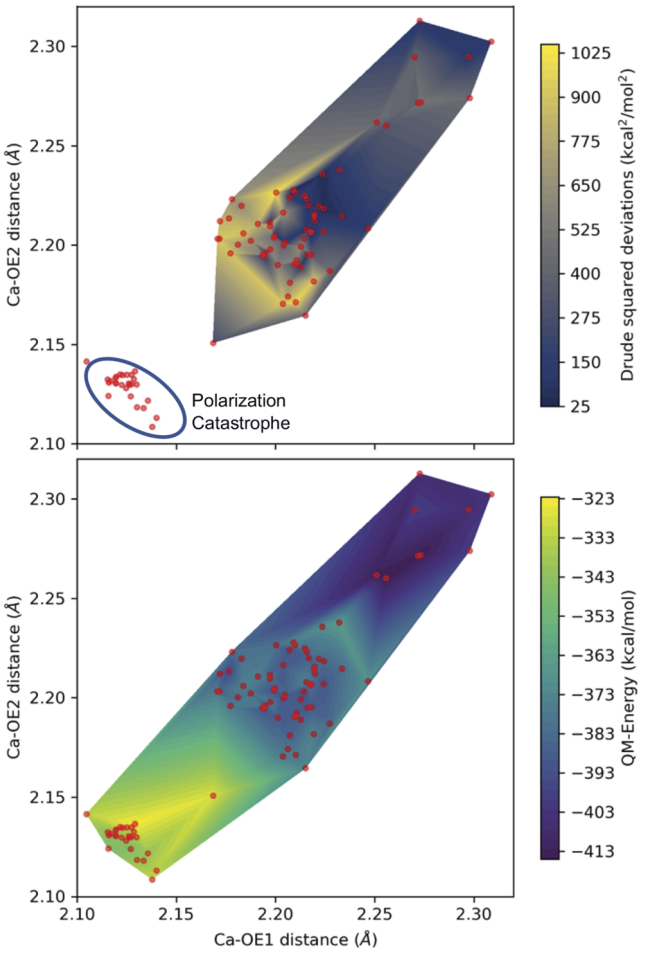


FIG. 7. (a) Drude FF squared energy deviations of Glu represented as functions of two collective variables—the distances in Å of Ca^{2+} to OE1 and OE2 carboxylate oxygen atoms of Glu. Red dots represent each of the Glu conformations as a function of these two collective variables. Colors represent the SD between the QM and MM energy. The region where polarization catastrophe occurs is circled and has SD values of ~10^{16}. (b) QM interaction energies for Ca^{2+}–dipeptide fragments as functions of the two collective variables described above for (a). Both surfaces are obtained by triangle-based linear interpolation of the data.

types. In the Drude FF, all four of these atoms are described by a single atom type, namely, OD2C2A, and, thus, have the same set of parameters. Note that Thole screening between this atom type and Ca^{2+} has not been implemented in the original FF, and Thole screening parameters were introduced only for ion–water oxygen interactions.[36] In Sec. III B, we show that its inclusion is vital to avoid over-polarization.

B. Reduction of parameter space and avoiding polarization catastrophe

The parameters that most significantly determine the interaction energies between a metal cation and an Asp-/Glu-dipeptide are the non-bonded LJ parameters ϵ, σ, between the carboxyl oxygen and the ion as well as the electrostatic forces between them. The partial charges had been carefully parameterized and are difficult to change due to their large degree of interdependency. The same is true for the polarizabilities and the Drude particle spring constants. However, the NBFIX option in CHARMM invokes pair-specific Thole screening factors t_{ij} and pair-specific LJ parameters σ_{ij}, which would ideally be driven and optimized against a panel of condensed matter simulations. It is important to note that the LJ parameters apply only to the nuclei, which are constrained to positions derived from QM reference structures; hence, optimization of LJ parameters will affect the total energy of the system but not the geometries of the Drude particles,[53,75] which do not experience any LJ potentials. Since the nuclei are constrained to the QM-optimized geometry, the only degrees of freedom during energy minimization are those of the Drude particle positions. Thus, including a pair-specific Thole screening factor will affect not only the energy of the system but also the locations of the Drude particles, although their

impact is relatively small except when there is a significant electron overlap.

To illustrate the effect of the pair-specific Thole parameter (t_{ij}) between the carboxylate oxygen and Ca^{2+}, we calculated the Drude-FF interaction energies between Ca^{2+} and Asp-/Glu-dipeptide for three different values of t_{ij} and compared them with QM interaction energies in Fig. 8. By default, if a pair-specific Thole is not specified for non-bonded pairs, $t_{ij} = \infty$ for that pair, i.e., $S_{ij} = 1$, and there is no electrostatic screening of the Coulomb potential. This is represented by the dashed line in Fig. 8. We also computed the interaction energies at $t_{ij} = 3.0$ and $t_{ij} = 2.6$, where $t_{ij} = 2.6$ results in a stronger electrostatic damping. Figure 8 illustrates the utility of the pair-specific damping factor in controlling polarization catastrophes in problematic conformations, without substantially altering the energy surface in the rest of the conformational space. For Glu [Fig. 8(a)], the catastrophe occurs in a large number of conformations, increasing the chances of it occurring in real simulations. For Asp [Fig. 8(b)], the catastrophe occurs in a much lower energy region; thus, it could be problematic even though only one conformation suffers from this phenomenon. Furthermore, even when the Thole parameter is introduced, if it is not strong enough ($t_{ij} = 3.0$), then some conformations can still have unrealistically low energies due to the tendency to overpolarize, but they are still of the same order of magnitude as the QM minimum energies. This may result in hard-to-detect over-polarization phenomena in simulations and hampers the development of balanced polarizable FFs.

C. Local environmental effects in backbone carbonyl–Ca^{2+} interactions

The dipeptide–cation dataset allows one to explicitly assess subtle, but important, effects of local changes in the electrostatic environment on the peptide–ion interactions. The chemical space mapped out in the current QM dataset is a good representative of the Ca^{2+}-binding sites found in the PDB surveys,[76,77] which show Asp/Glu carboxylates to be the most frequent first-shell ligands followed by the backbone carbonyl and water ligands. In accord with the trends found in the PDB surveys, the most common atoms that coordinate Ca^{2+} are the carboxylate oxygens for the Asp-/Glu-dipeptide (OE1, OE2, OD1, and OD2) and the acetylated terminal carbonyl oxygen (OY) or backbone carbonyl oxygen (O) for the other dipeptides, as shown in Table SI 1. Therefore, the dataset enables the exploration of the potential impact of the local changes in the chemical environment on the peptide–Ca^{2+} interactions.

In the Drude protein FF, OY and O are represented by the same atom type (OD2C1A), which, while making the parameter exploration easier, may reduce the accuracy in the description of ion–ligand interactions. Indeed, the SD between the MM and QM interaction energies as a function of Ca^{2+}–OY and Ca^{2+}–O distances in FigG. SI 2 indicates a slight asymmetry in interaction energies resulting from the acetylation and an increase in polarity of the coordinating carbonyl oxygen, which is not captured in the FF if the same

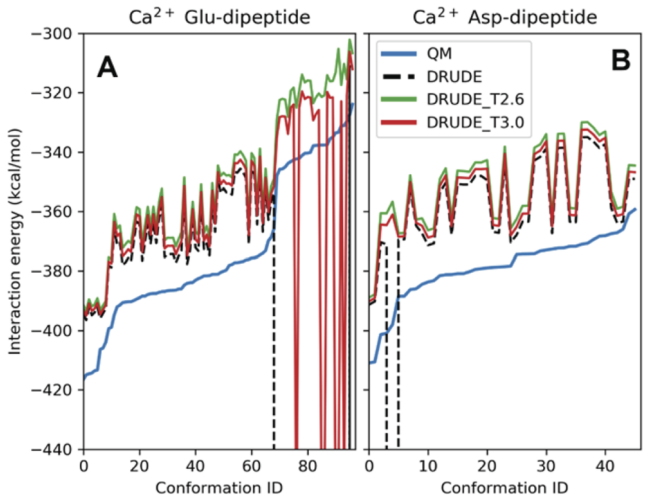


FIG. 8. Interaction energies of the Drude FF compared with QM for three values of the pair-specific Thole parameter for interaction between Ca^{2+} (CALD) and Glu/Asp carboxylate oxygen (OD2C2A). The values of the Thole parameter are $t_{ij} = \infty$ (*dashed line*), 2.6 (*green*), and 3.0 (*red*), which are illustrated for modeling interaction energy between Ca^{2+} and Glu (a)- and Asp-dipeptide (b), respectively.

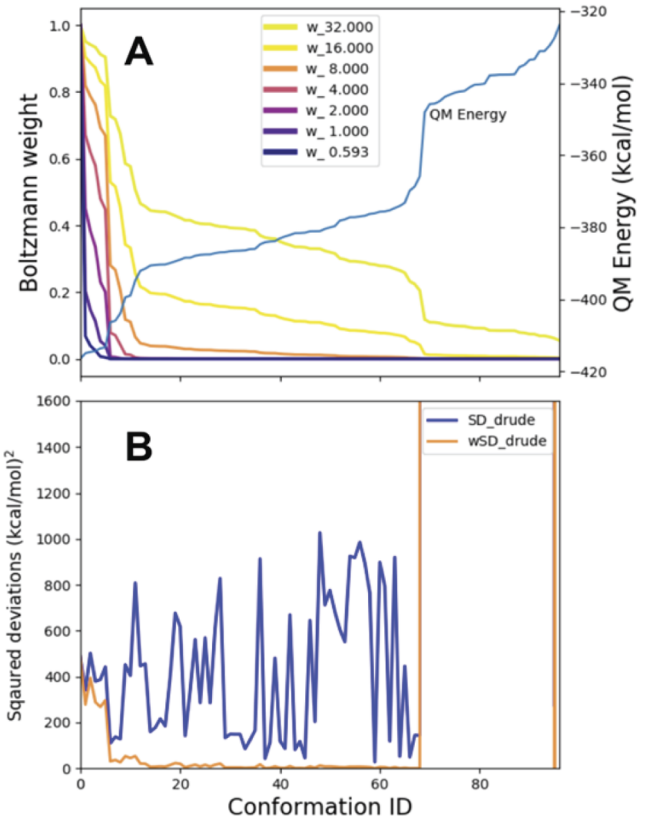


FIG. 9. Boltzmann weights applied to Glu-dipeptide–Ca^{2+} interactions. (a) Boltzmann weight vs conformation ID at various RTs (0.593–32). The blue curve is the corresponding reference QM interaction energy. (b) Boltzmann-weighted squared deviations with RT = 8 (wSD) and unweighted squared deviations (SDs) plotted for the Drude FF interaction energies.

TABLE I. Parameter change summary for the pair-specific Ca-OD2C2A Thole parameter (t_{ij}) and LJ parameter (σ_{ij}).[a]

Parameter	DRUDE	DRUDE_T2.6	DRUDE-wRMSD
t_{ij} (NBTHOLE)	N/A	2.600 00	1.400 00
σ_{ij}(NBFIX) (Å)	3.515 00	3.515 00	2.891 43
RMSD (Asp) (kcal/mol)	2.28×10^7	28.93	8.43
RMSD (Glu) (kcal/mol)	8.17×10^7	24.00	12.99

[a]The last column lists the parameters of DRUDE_OPTIMUM also illustrated by Fig. SI 3.

atom type is used for both OY and O. The deviations from QM calculated interaction energies generally occur when the Ca^{2+}–OY and Ca^{2+}–O distances are between 2.10 Å and 2.25 Å, where a significant electronic overlap (repulsion) exists.

D. Optimizing parameters against DFT energies using a Boltzmann-weighted RMSD

It is crucial to consider carefully how to evaluate the relationships between energy surfaces represented in MM models and the DFT-based energy surfaces. Since it is not possible to fit all parts of the two surfaces to arbitrary precision, which parts of the surfaces are most important? A common and very successful approach is to focus on a set of selected interaction directions and meticulously scan them using resulting QM data to fit the function. One fitting criterion often used is the RMSD between the two surfaces defined in Eq. (4). This method puts more weight on parts of the energy surface whose absolute values are larger. However, the weight on the minima may not be enough. The true weight of each ion position should closely represent the Boltzmann weight of the system at those positions. One way to account for this is to have a higher density of reference structures near the minima, with the number of grid points for sampling being proportional to the Boltzmann weight. This could be more expensive, depending on the number of points. Another approach is to take a grid of points, calculate the Boltzmann weights *a posteriori*, and apply them to the fitting function. Taking this approach yields an adjusted scoring function, the Boltzmann-weighted RMSD (wRMSD),

$$wRMSD = \left[\sum_i w_i \left(E^i_{QM} - E^i_{MM}\right)^2\right]^{\frac{1}{2}}, \tag{9}$$

where we have modified the RMSD in Eq. (4) by including a Boltzmann factor,

$$w_i = A \exp\left[\frac{-E^i_{QM}}{RT}\right], \tag{10}$$

where A is the normalization constant (so that $\sum_i w_i = 1$) And RT is the "temperature factor" that does not has any physical meaning, but affects the flatness of the distribution.

Figure 9 shows an example of applying Boltzmann weights to the Glu-dipeptide–Ca^{2+} system. Figure 9(a) shows how the Boltzmann weights (w_i) vary with increasing RT. The higher the QM interaction energy, the less the weight, but the temperature factor (RT) determines the degree of relative importance of the lower energy conformations. $RT = \infty$ is the same as using the RMSD since all weights would be identical, whereas low values of RT will put more relative weight on the minima. Figure 9(b) shows how the weighted squared deviation differs from the unweighted one for RT = 8. While the weighted squared deviations generally put more emphasis on low-energy conformations near the QM minima, it does blow up for conformations where polarization catastrophes occur. Thus, with an appropriate choice of RT, one can get a scoring function for the parameter optimization that puts more weight on the low-energy minima, but can still detect large outliers at other energies.

Supplementary material, Table S1, shows that in the majority of Glu- and Asp-dipeptide conformations, the nearest atoms to Ca^{2+} are OE1, OE2, OD1, and OD2, which are given as a single atom type: OD2C2A. This means that they are identical in their non-bonded interaction with ions. Thus, to optimize the interactions of Ca^{2+} with carboxylate-containing dipeptides, we targeted the pair-specific

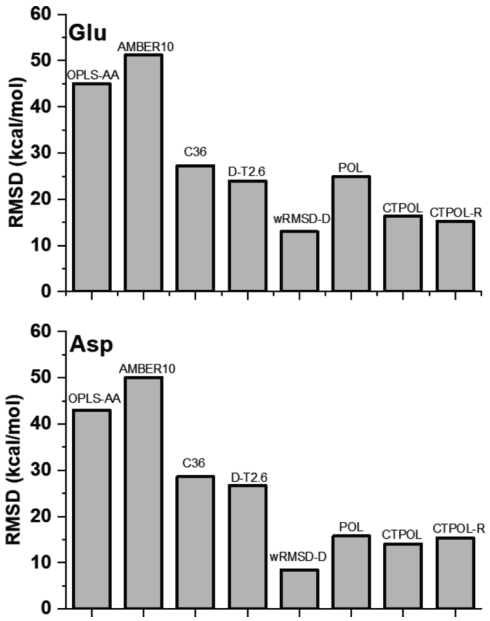


FIG. 10. RMSD estimated with Eq. (4) for the ensemble of conformations of Glu:Ca^{2+} (top panel) and Asp:Ca^{2+} (bottom panel) for various FFs used in our study. Drude data are shown for the Thole parameter set to 2.6 (D-T2.6) and optimized LJ and Thole parameters using Boltzmann-weighted RMSD (wRMSD-D). CTPOL parameters were fitted for the AMBER10 FF with (i) only a local polarization response term (POL), (ii) unrestricted charge-transfer contribution (CTPOL), and (iii) with restricted charge-transfer contribution (CTPOL-R).

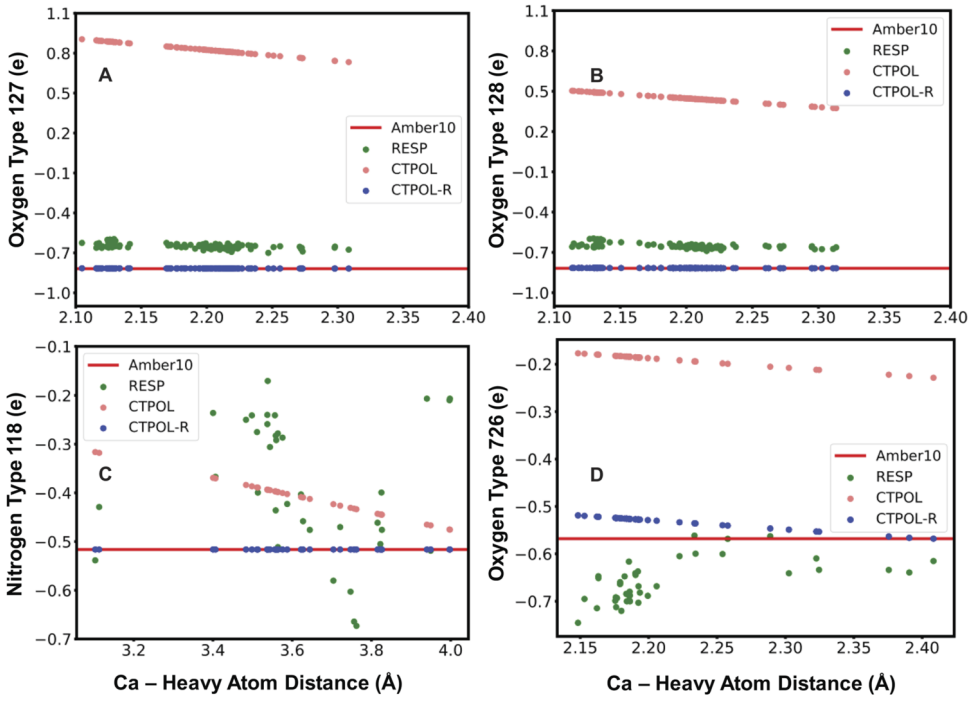


FIG. 11. Partial charge of ligand atoms in Ca^{2+}–Glu-dipeptide vs distance between an atom and Ca^{2+}. (a) and (b) show charges for the carboxylate oxygens (atom-types 127 and 128, as shown in Fig. SI 1). (c) and (d) show charge vs distance dependence for backbone nitrogen (atom-type 118) and backbone oxygen (atom-type 726), respectively. All charges are shown in electron units. The red line represents the atom's partial charge in the standard AMBER10 FF. The calculated RESP partial charges are shown for comparison as green dots.

interaction between the Ca^{2+} (i) and OD2C2A (j) and optimized the Thole parameter t_{ij} (NBTHOLE) and the LJ σ_{ij} parameter (NBFIX). This was done by running Drude interaction energy calculations for eight different NBTHOLE values, each with eight different NBFIX values, resulting in 64 different parameter combinations. The NBTHOLE values ranged from 1.2 to 2.6, whereas the NBFIX values ranged from 2.72 to 3.92. For each of these parameter combinations, the wRMSD given by Eq. (9) was calculated for Glu- and Asp-dipeptide interaction energies, and the global minimum with respect to the wRMSD was chosen as the optimum parameter set. The changes in parameters are summarized in Table I.

Figure SI 3 shows the improvement in accuracy due to the new parameter set over the original one. The optimized Drude FF (referred to as Drude-wRMSD) no longer displays any polarization catastrophe and gives a much closer fit to the QM interaction energies, particularly near the minima. It is apparent from Fig. SI 3, Fig. 8, and Table I that electrostatic optimization via the Thole parameter alone cannot reproduce QM energies. The LJ σ_{ij}-parameter also has to be changed in order to match QM energy across a broader range of conformations. Optimization of Thole and LJ parameters has not only produced better RMSDs between QM and MM energies but also reduced the fluctuations in the energy trend. This implies that the ranking of conformations by energy more closely resembles the ranking of QM energies for most of the conformational space. However, the plateau region of the QM energy of Glu-dipeptide, which is present in other FFs (see Fig. 6) as well as in

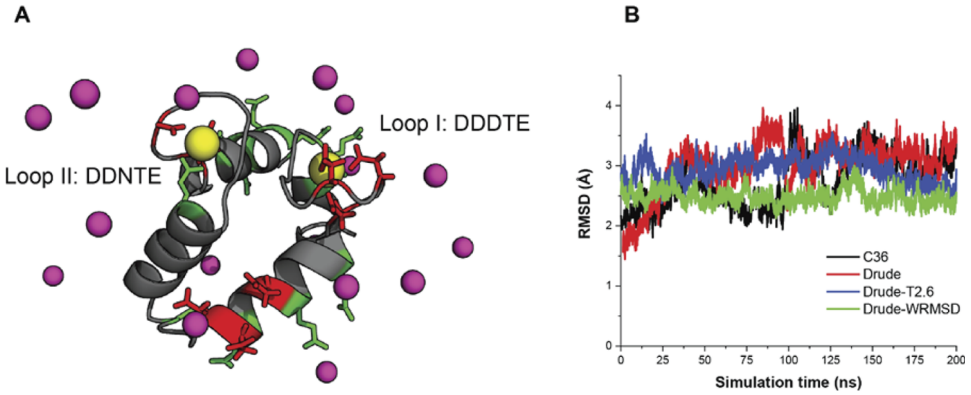


FIG. 12. (a) The N-lobe of the CaM protein with Loop I and II with two bound Ca^{2+} ions (gold spheres). The positions of the aspartate residues are shown in red sticks, while glutamates are shown in green sticks. The Ca^{2+} ions from the bulk solution are shown as magenta spheres. Water molecules and Cl^- ions are not shown for clarity. (b) Time traces of the RMSD for protein heavy atom coordinates relative to the x-ray structure (PDBID:1CLL) for the C36 and Drude FFs. First 50 ns of all MD runs were discarded, and only production runs of 200 ns were used.

Drude_T2.6, is absent for this parameter set. The weighted RMSD function puts a very low weight on this part of the conformational space due to the high energies. However, this is an important part of the conformational space as it comprises the conformations with the shortest distances between Ca^{2+} and the carboxylate oxygens. A larger exploration of the parameter space may be required to remedy the discrepancy in this region, and the scoring function may also need to be revisited in order to treat these regions on a more equal footing.

E. The extent of charge transfer in cation interactions with carboxylate groups

Although future development of a balanced Drude FF for cation–protein interactions is under way, it may still be limited when charge-transfer effects between a cation and coordinating ligands are significant. We have previously performed a PDB survey to elucidate the effect of the secondary shell ligands on cation-binding to metalloproteins using DFT and DFT-tight binding (TB) methods.[11,76] We found that perturbation of the charges on coordinating ligands due to Ca^{2+}-binding is significant, amounting up to 15%–20% of partial-charge change on the coordinating oxygen atoms. This effect is not limited to ligands in the first coordination shell, but impacts ligands in the second shell, albeit to a lesser extent. Since the CTPOL formalism[41,42] incorporates both local polarization (POL) and charge-transfer (CT) effects into the interaction energy (see Sec. II F), we employed this FF model to study partial-charge transfer for the two challenging Asp–Ca^{2+} and Glu–Ca^{2+} systems to potentially present a strategy for FF re-calibration of cation–peptide interactions. Importantly, it allows one to investigate a model containing just a local polarization response term (POL) or a model that additionally includes the charge-transfer contribution (CTPOL).

Table SI 2 summarizes the parameters in the CTPOL FF used, namely, atomic polarizabilities and charge transfer a_L and b_L coefficients in Eq. (6) fitted for the AMBER10 FF (see Sec. II). The coefficients in Eq. (6) used QM interaction energies as the input. It is important to note that the choice of QM level of theory for the reference dataset affects the absolute values of the total energies. However, Ngo *et al.*[11] studied different all-electron DFT functionals and showed that the *absolute* binding energies computed using different functionals and basis sets can vary by up to 10% depending on the method, but the corresponding *relative* binding energies vary by only 4%–5% relative to calculations performed with higher basis sets. Hence, our study will focus on the trends and elucidate areas to pay attention to in metalloprotein FF development.

The RMSD values in Fig. 10 demonstrate the apparent usability of Drude FFs with a control for polarization catastrophes via a carefully developed set of NBFIX/Thole parameters for simulating larger systems. It also shows that a standard force field (in this case Amber10) extended by a local polarization term [Eq. (8), POL] can be optimized against available higher-level data. This POL FF significantly improves the performance of the original FF. Adding a charge transfer term (CTPOL) without any constraints on the charge transfer extent led to further improvement in the RMSD, as evident in Fig. 10(b). However, the charge transfer parameters in Eq. (6) yield unphysical partial charges such as a negative charge on Ca^{2+}, probably because they were determined to reproduce the relative QM interaction energies as the objective function without any constraints on the amount of charge transfer. Hence, they compensate for the inherent errors of the standard AMBER10 FF, which yields a RMSD from QM energies that is generally greater than that of C36 (Fig. 5). One way to address this issue is to restrict the amount of charge transfer in the model denoted as CTPOL-R. Figure 11 shows how this improves the charges on a few selected atoms, particularly the carboxylate oxygens of Glu-dipeptide [Figs. 11(a), 11(b), and 11(d)].

However, implementing this fix alone leads to an RMSD of 35.7 kcal/mol, which is clearly not satisfactory. If, however, we re-optimize the original AMBER10 FF vdW parameters of atoms involved in direct interactions with Ca^{2+}, we obtain a reasonable RMSD of 15.4 kcal/mol for CTPOL-R (CTPOL with restricted charge transfer), which is comparable to the RMSD of CTPOL without any restriction (16.4 kcal/mol). The list of adjusted vdW parameters is provided in supplementary material, Table SI 3. Although, the resulting charge transfer term is only about 2 kcal/mol, tweaking the original parameters of the AMBER10 FF was crucial for simultaneously correcting the signs of the charge transfer and reducing the RMSD.

F. Evaluation of Drude-FF parameters in metalloprotein simulations

The Drude, Drude_T2.6, and Drude-wRMSD polarizable FF parameters were assessed and compared with the C36 parameters by using them in MD simulations of the N-lobe of the human calmodulin (CaM) protein shown in Fig. 12(a). The RMSD values for all FFs collected in Fig. 12(b) are comparable, with significant flexibility observed for loops I and II (RMSD ~ 2.4 Å–2.7 Å). The highest RMSD values are observed for the truncated portion of the central helix and are related to partial bending and unwinding (region-specific RMSD > 3.5 Å). While similar dynamics has been reported for the central helix in nuclear magnetic resonance (NMR), spectroscopic and modeling studies,[18,78–80] it may still be driven by the choice of the reduced model.

Table II compares the coordination numbers of Ca^{2+} in Loop I and Loop II binding sites obtained with different FFs and the

TABLE II. Calcium coordination numbers for EF-hand Loop I and Loop II sites.[a]

	Excluding water (including water)	
	Loop I	Loop II
C36	6.45 (7.31)	6.71 (7.64)
Drude	5.95 (7.31)	5.55 (6.45)
Drude-T2.6	5.84 (7.01)	5.81 (6.91)
Drude-wRMSD	5.99 (5.99)	5.95 (5.97)
ECCR[74]	5.94 (7.00)	7.03 (7.04)
1CLL[60]	6.00 (7.00)	6.00 (7.00)

[a]Ca^{2+} coordination numbers were determined by integration over the first peak of the ion-oxygen RDF, which included oxygen atoms of water and amino-acid residues in the RDF calculations; numbers with parentheses include both protein and water ligands, whereas those without exclude the water ligand.

scaled-charge approach referred to as ECCR[17,18] with those found in the x-ray structure. The high-resolution (1.7 Å) crystal structure (PDBID:1CLL) with the positions of the crystallographic waters resolved shows six protein ligands and a water molecule directly heptacoordinated to the divalent cation in each site. An important and unique feature of the binding sites reported in 1CLL is the presence of bidentate (Glu) and monodentate coordination (Asp) modes.[65] The reproduction of the monodentate coordination by aspartates present in the x-ray structure represents a significant challenge for the additive FFs, where partial charge is distributed equally between two coordinating oxygens.[18,81] Indeed, the C36 FF exhibits a shift from monodentate to predominantly bidentate coordination for the binding site in loop II (Asp 56) and to a lesser extent in loop I (Asp 22), resulting in an average of more than six calmodulin oxygens in the calcium coordination shell.

Interestingly, the scaling approach used in the ECCR study with the charge on Ca^{2+} scaled down to +1.5*e* still led to bidentate coordination in Loop II. The authors suggested that the coordination number of 7 observed in the charge-scaling approach is due to the recruitment of an additional aspartate (Asp 64) into the ion coordination.[18] In contrast, we did not observe stable coordination by Asp 64 in any of our simulations. All of the Drude models resulted in protein coordination numbers between 5.55 (Drude) and 5.98 (Drude-wRMSD), showing predominantly monodentate coordination for both aspartates (Asp 22 and Asp 56) in accord with the coordination mode reported in the x-ray structure.[18] The Drude-wRMSD model shows a near ideal coordination mode for the protein ligands, but fails to reproduce the retention of a Ca^{2+}-bound water molecule in Loop I and Loop II.

Analysis of minimal distances between Ca^{2+} and protein ligands reveals a potential issue that may explain why the Drude-wRMSD model resulted in the release of a water molecule from the first coordination sphere. The Drude-wRMSD model routinely shows unphysical distances <1.5 Å between the cation and

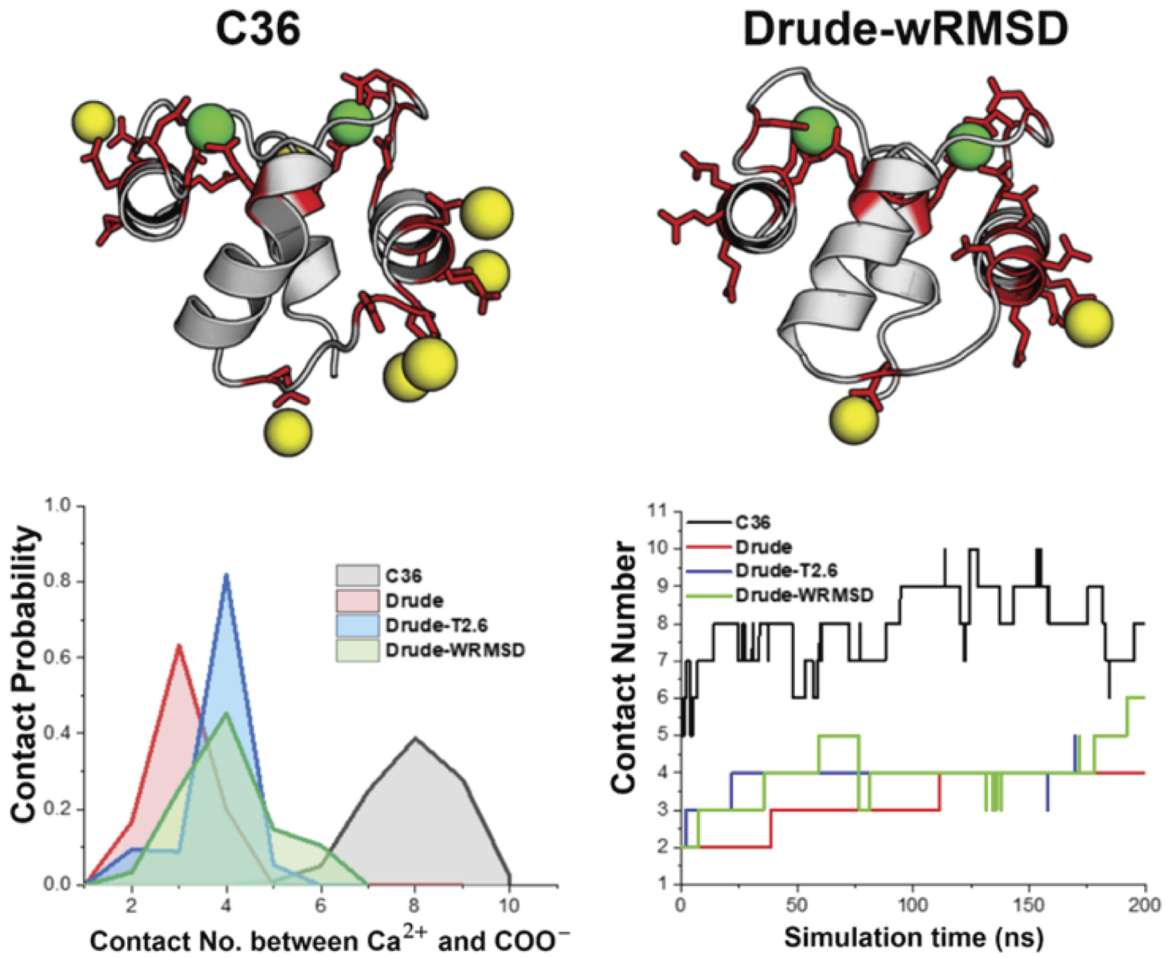


FIG. 13. (Top) Characteristic snapshots of a single EF-hand CaM in the solution with 150 mM $CaCl_2$ with contact numbers 9 and 4 observed in simulations with C36 (left) and Drude-wRMSD (right) FFs, respectively. In both snapshots, the two Ca^{2+} ions bound to sites in Loop I and II are shown as green spheres, while Ca^{2+} ions recruited from the bulk solutions are shown as gold spheres. Asp and Glu residues of CaM are shown as red sticks. Bottom left: The distribution of the contacts between Ca^{2+} and COO^{-} reported for all FFs considered in this work. Bottom right): Time traces for CN calculated with different FFs.

negatively charged lone pairs located on the carbonyl or carboxylate oxygens, both for ions bound to sites in Loop I and Loop II and those recruited from the bulk phase to coordinate solvent-exposed residues. This issue has also been noted in all the simulations performed with the original Drude parameters for Ca^{2+} with ion–oxygen lone-pair distances as short as 1.55 Å, the corresponding Ca^{2+}–O(carboxylate) coordinating distances between 1.9 Å and 2.2 Å, and the first peak of the radial distribution function (RDF) located at 2.05 Å, which is significantly shorter than 2.30 Å–2.45 Å observed in other proteins.[74] The first peak in the RDF between Ca^{2+} present in the *bulk* solution and carboxylate oxygens is located at 2.10 Å for Drude-wRMSD and at 2.35 Å for Drude. The introduction of the Thole parameter equal to 2.6 combined with adjusted NBFIX values appears to correct this issue. Using the Drude-T2.6 FF, the shortest ion–lone-pair distance is 1.83 Å, the average distance is 1.96 Å, and the first peak in the RDF between Ca^{2+} and the carboxylate oxygens is located at 2.40 Å, in accord with the results of the PDB surveys.[74,76] The unphysically short ion coordination distances observed in Drude-wRMSD lead to an "over-stabilization" phenomenon and presumably an over-binding on the protein surface, as suggested by studies of ion transport in ryanodine receptors[37] and porins.[39]

Recent comparative analysis of MD simulations and capillary electrophoresis experiments for dications binding to insulin[17,19] indicates that specific and very tight binding of cations in the physiological pH range leads to over-accumulation of mobile charges on the protein surface modeled with non-polarizable FFs.[17] For example, up to 20 Ca^{2+} were reported to bind stably to the full CaM structure, in stark contrast with the anticipated four ions bound to sites present in the EF hands.[17] By introducing ECCR corrections with CHARMM36 parameters, the overall number of Ca^{2+} ions reduced drastically to ~6, or 3 cations per lobe. To compare the performance of the polarizable FFs considered in our study to results reported by Duboué-Dijon *et al.*,[17] we computed probability distributions for the Ca^{2+}–carboxylate contact number (CN). The contact distance R between a cation and an Asp/Glu carboxylate group was defined based on the position of the first minimum in the RDF between Ca^{2+} and the carboxylate carbon atom. It was set to R = 4.1 Å in accord with R = 4.0 Å used by Duboué-Dijon *et al.*[19] While our simulation results obtained for a single CaM lobe containing two sites in Loop I and II cannot be directly compared with those obtained by Duboué-Dijon *et al.*[17,19] for the *full* CaM structure, the overall trend appears to be similar. In simulations using the C36 FF, the average CN is 8.5 with a single carboxylate group coordinating up to two cations for up to hundreds of nanoseconds (see Fig. 13). Compared to C36, the average CN of 4 for all the Drude models studied is much smaller. However, the Drude-wRMSD features a very broad distribution with CN up to 6 routinely present.

The charge scaling used by Duboué-Dijon *et al.*[17,19] led to the destabilization of the cation-binding sites present in the EF hands (Loop I and Loop II sites) in under 60 ns of production MD runs. In our simulations, however, no cation unbinding from Loop I and Loop II was observed in 200 ns. We used 150 mM $CaCl_2$, which is expected to increase the cation concentration at the protein surface. In simulations performed with Drude parameters, no cation exchanges were observed; e.g., once Ca^{2+} is recruited from the bulk solution to the binding pocket, it remained bound for the whole duration of the simulation. This is especially apparent with the original Drude force field, where CN gradually rose from 2 to 3 and then to four cations stably bound to the carboxylate residues facing the solution (Fig. 13). For the Drude-wRMSD simulations, the cations bound to Loop I and Loop II remain coordinated by protein atoms only, and no water molecule was recruited to the first coordination shell.

IV. CONCLUSIONS AND OUTLOOK

In summary, we have performed a comprehensive benchmarking of the existing FFs for Ca^{2+}–dipeptide interaction energies against a comprehensive QM dataset. Several areas for the potential improvement of metalloprotein models in the context of the polarizable FFs were identified, notably, undamped electrostatic forces causing the Drude oxygen to overlap with Ca^{2+} [Fig. 6(b)] when it is near the Asp/Glu carboxylate. We show how this may be ameliorated by an illustrative parameterization of Ca^{2+} interaction energies with Glu/Asp-dipeptides using RMSD and weighted RMSD approaches. This leads to a better performance for the reproduction of the gas-phase energetics with some notable exceptions present in the broad conformational space sampled in the QM dataset. With the CTPOL method, problems related mainly to unphysical charges on Ca^{2+} arose in parameterizing the same Glu/Asp-dipeptide in a similar region of the conformational space. This was substantially remedied by imposing restrictions on the amount of charge transfer and reparameterizing some of the original parameters of the additive FF. However, none of the parameter sets tested in our study are at a stage where they can be recommended for large-scale metalloprotein simulations in the condensed phase.

We have taken a first step toward relating the parameter space to the conformational space with the current analysis. By expressing the conformational space in terms of distances between the cation and coordinating atoms, we may determine better parameter subspaces using RMSD and wRMSD of interaction energies for fitting. The next logical goal would be to test other scoring functions such as binding energies rather than interaction energies, or relative interaction energies instead of absolute interaction energies. Each set of optimized parameters obtained for a subset of the dataset and parameter space should be tested by performing condensed matter simulations. This would allow us to identify the strategy that produces the best relative improvement, which can then be applied to the whole dataset with a larger parameter space. The combination of QM-led initial parameter development and comprehensive testing in the condensed phase would help us to capture more accurate dynamical and structural properties of ion binding to biomolecules.

While a comprehensive QM dataset complements sparse experimental data and helps us to elucidate the key problems in parameterization, certain gas-phase QM conformations may not be pertinent in a solvated protein environment where the effective dielectric constant is generally >1. In the future, we advocate the use of micro-solvated QM systems such as metal-bound dipeptides surrounded by nearby water molecules or larger solvated QM/MM systems with potential applications of force-matching algorithms.[20] Another avenue for future work would be to derive CTPOL parameters and systematically optimize the original FF parameters to reproduce micro-solvated QM or solvated QM/MM data as well as available experimental hydration structures and relative hydration free

energies of *all* cations of the same charge.[82] Such a parameterization approach would lead to force fields that can better reproduce the complex environments of biologically important metalloproteins containing more than one type of cation.

SUPPLEMENTARY MATERIAL

See the supplementary material for atom-type definitions in various FFs, parameters for the CTPOL-R/AMBER10 model, and figures for additional energy scans.

ACKNOWLEDGMENTS

The work in SYN and DRS labs was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (Discovery Grant No. RGPIN-315019 to SYN and Discovery Grant No. RGPIN-2019-03976 to DRS). C.L. thanks Academia Sinica (Grant No. AS-IA-107-L03) and the Ministry of Science and Technology, Taiwan (Grant No. MOST-98-2113-M-001-011), for support. K.S.A. is supported by the University of Calgary Provost Doctoral Fellowship. X.H. is grateful for a doctoral fellowship by the China Scholarship Council. The calculations for this submission were enabled by funding from the NSERC-RTI program used to acquire the CPU–GPU cluster www.glados.ucalgary.ca and by the Resource Allocation Award from Compute Canada. Panel A molecular graphics was prepared using python package NGLView. Figure 6 of Ref. 83 has been prepared with python package NGLView.

AUTHORS' CONTRIBUTIONS

K.S. Amin and X. Hu contributed equally to this work.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] D. J. Huggins, P. C. Biggin, M. A. Dämgen, J. W. Essex, S. A. Harris, R. H. Henchman, S. Khalid, A. Kuzmanic, C. A. Laughton, J. Michel, A. J. Mulholland, E. Rosta, M. S. P. Sansom, and M. W. van der Kamp, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **9**(3), e1393 (2019).
[2] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, Annu. Rev. Biophys. **41**, 429–452 (2012).
[3] J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell, Chem. Rev. **116**(9), 4983–5013 (2016).
[4] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, Jr., Nat. Methods **14**(1), 71–73 (2017).
[5] E. Flood, C. Boiteux, B. Lev, I. Vorobyov, and T. W. Allen, Chem. Rev. **119**(13), 7737–7832 (2019).
[6] Z. F. Jing, C. W. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J. P. Piquemal, and P. Y. Ren, Annu. Rev. Biophys. **48**, 371–394 (2019).
[7] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **3**(2), 198–210 (2013).
[8] M. M. Reif, P. H. Hünenberger, and C. Oostenbrink, J. Chem. Theory Comput. **8**(10), 3705–3723 (2012).
[9] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, J. Phys. Chem. B **105**(28), 6474–6487 (2001).
[10] H. Li, V. Ngo, M. C. Da Silva, D. R. Salahub, K. Callahan, B. Roux, and S. Y. Noskov, J. Phys. Chem. B **119**, 9401–9416 (2015).
[11] V. Ngo, M. C. da Silva, M. Kubillus, H. Li, B. Roux, M. Elstner, Q. Cui, D. R. Salahub, and S. Y. Noskov, J. Chem. Theory Comput. **11**(10), 4992–5001 (2015).
[12] X. D. Peng, Y. B. Zhang, H. Y. Chu, Y. Li, D. L. Zhang, L. R. Cao, and G. H. Li, J. Chem. Theory Comput. **12**(6), 2973–2982 (2016).
[13] H. MacDermott-Opeskin, C. A. McDevitt, and M. L. O'Mara, J. Chem. Theory Comput. **16**(3), 1913–1923 (2020).
[14] J. Yoo and A. Aksimentiev, Phys. Chem. Chem. Phys. **20**(13), 8432–8449 (2018).
[15] T. Dudev and C. Lim, Chem. Rev. **114**(1), 538–556 (2014).
[16] P. Li and K. M. Merz, Chem. Rev. **117**(3), 1564–1686 (2017).
[17] E. Duboué-Dijon, M. Javanainen, P. Delcroix, P. Jungwirth, and H. Martinez-Seara, J. Chem. Phys. **153**(5), 050901 (2020).
[18] M. Kohagen, M. Lepšík, and P. Jungwirth, J. Phys. Chem. Lett. **5**(22), 3964–3969 (2014).
[19] E. Duboué-Dijon, P. Delcroix, H. Martinez-Seara, J. Hladílková, P. Coufal, T. Křížek, and P. Jungwirth, J. Phys. Chem. B **122**(21), 5640–5648 (2018).
[20] O. Akin-Ojo, Y. Song, and F. Wang, J. Chem. Phys. **129**(6), 064108 (2008).
[21] J. C. Li and F. Wang, J. Chem. Phys. **143**(21), 074311 (2015).
[22] P. Li, L. F. Song, and K. M. Merz, Jr., J. Chem. Theory Comput. **11**(4), 1645–1657 (2015).
[23] P. Li, L. F. Song, and K. M. Merz, Jr., J. Phys. Chem. B **119**(3), 883–895 (2015).
[24] J.-P. Piquemal, H. Chevreau, and N. Gresh, J. Chem. Theory Comput. **3**(3), 824–837 (2007).
[25] S. W. Rick, S. J. Stuart, and B. J. Berne, J. Chem. Phys. **101**(7), 6141–6156 (1994).
[26] H. A. Stern, G. A. Kaminski, J. L. Banks, R. Zhou, B. J. Berne, and R. A. Friesner, J. Phys. Chem. B **103**(22), 4730–4737 (1999).
[27] Y. Luo, W. Jiang, H. B. Yu, A. D. MacKerell, and B. Roux, Faraday Discuss. **160**, 135–149 (2013).
[28] J.-P. Piquemal, L. Perera, G. A. Cisneros, P. Ren, L. Pedersen, and T. A. Darden, J. Chem. Phys. **125**(5), 054511 (2006).
[29] 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, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, J. Phys. Chem. B **114**(8), 2549–2564 (2010).
[30] N. Manin, M. C. da Silva, I. Zdravkovic, O. Eliseeva, A. Dyshin, O. Yaşar, D. R. Salahub, A. M. Kolker, M. G. Kiselev, and S. Y. Noskov, Phys. Chem. Chem. Phys. **18**(5), 4191–4200 (2016).
[31] A. V. Aleksandrov, B. Roux, and A. D. MacKerell, J. Chem. Theory Comput. **16**, 4655 (2020).
[32] S. Patel and C. L. Brooks III, J. Comput. Chem. **25**(1), 1–16 (2004).
[33] Z.-Z. Yang, J.-J. Wang, and D.-X. Zhao, J. Comput. Chem. **35**(23), 1690–1706 (2014).
[34] T. Dudev, M. Devereux, M. Meuwly, C. Lim, J.-P. Piquemal, and N. Gresh, J. Comput. Chem. **36**(5), 285–302 (2015).
[35] Z. Jing, C. Liu, R. Qi, and P. Ren, Proc. Natl. Acad. Sci. U. S. A. **115**(32), E7495–E7501 (2018).
[36] H. B. Yu, T. W. Whitfield, E. Harder, G. Lamoureux, I. Vorobyov, V. M. Anisimov, A. D. MacKerell, and B. Roux, J. Chem. Theory Comput. **6**(3), 774–786 (2010).
[37] A. Zhang, H. Yu, C. Liu, and C. Song, Nat. Commun. **11**(1), 922 (2020).
[38] V. Ngo, J. K. Fanning, and S. Y. Noskov, Adv. Theory Simul. **2**(2), 1800106 (2019).
[39] J. D. Prajapati, C. Mele, M. A. Aksoyoglu, M. Winterhalter, and U. Kleinekathöfer, J. Chem. Inf. Model. **60**(6), 3188–3203 (2020).
[40] F. Villa, A. D. MacKerell, B. Roux, and T. Simonson, J. Phys. Chem. A **122**(29), 6147–6155 (2018).
[41] D. V. Sakharov and C. Lim, J. Comput. Chem. **30**(2), 191–202 (2009).
[42] D. V. Sakharov and C. Lim, J. Am. Chem. Soc. **127**(13), 4921–4929 (2005).
[43] M. Ropo, M. Schneider, C. Baldauf, and V. Blum, Sci. Data **3**(1), 160009 (2016).
[44] M. Ropo, V. Blum, and C. Baldauf, Sci. Rep. **6**(1), 35772 (2016).
[45] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**(18), 3865–3868 (1996).

[46]M. Schneider and C. Baldauf, arXiv:1810.10596 (2018).
[47]A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**(7), 073005 (2009).
[48]V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**(11), 2175–2196 (2009).
[49]X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, New J. Phys. **14**(5), 053020 (2012).
[50]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**(1), 141–151 (1999).
[51] J. A. Hartigan and M. A. Wong, J. R. Stat. Soc., Ser. C **28**(1), 100–108 (1979).
[52]C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, J. Phys. Chem. **97**(40), 10269–10280 (1993).
[53]F.-Y. Lin, J. Huang, P. Pandey, C. Rupakheti, J. Li, B. T. Roux, and A. D. MacKerell, Jr., J. Chem. Theory Comput. **16**(5), 3221–3239 (2020).
[54]V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**(3), 712–725 (2006).
[55]P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, PLoS Comput. Biol. **13**(7), e1005659 (2017).
[56]S. Jo, T. Kim, V. G. Iyer, and W. Im, J. Comput. Chem. **29**(11), 1859–1865 (2008).
[57]G. Lamoureux and B. Roux, J. Chem. Phys. **119**(6), 3025–3039 (2003).
[58]B. T. Thole, Chem. Phys. **59**(3), 341–350 (1981).
[59]E. Harder, V. M. Anisimov, T. W. Whitfield, A. D. MacKerell, and B. Roux, J. Phys. Chem. B **112**(11), 3509–3521 (2008).
[60]T. J. Giese and D. M. York, J. Chem. Phys. **120**(21), 9903–9906 (2004).
[61] C. W. Liu, R. Qi, Q. T. Wang, J. P. Piquemal, and P. Y. Ren, J. Chem. Theory Comput. **13**(6), 2751–2761 (2017).
[62]J. Kennedy and R. Eberhart, paper presented at the Proceedings of ICNN'95 - International Conference on Neural Networks, 1995.
[63]R. Poli, J. Kennedy, and T. Blackwell, Swarm Intell. **1**(1), 33–57 (2007).
[64]A. Lee, https://pythonhosted.org/pyswarm/, 2014.
[65]R. Chattopadhyaya, W. E. Meador, A. R. Means, and F. A. Quiocho, J. Mol. Biol. **228**(4), 1177–1192 (1992).
[66]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**(2), 926–935 (1983).
[67]J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid, J. Chem. Phys. **153**(4), 044130 (2020).
[68]T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**(12), 10089–10092 (1993).
[69]J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**(3), 327–341 (1977).
[70]S. Jo, X. Cheng, J. Lee, S. Kim, S.-J. Park, D. S. Patel, A. H. Beaven, K. I. Lee, H. Rui, S. Park, H. S. Lee, B. Roux, A. D. MacKerell, Jr., J. B. Klauda, Y. Qi, and W. Im, J. Comput. Chem. **38**(15), 1114–1124 (2017).
[71] P. E. M. Lopes, J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. MacKerell, J. Chem. Theory Comput. **9**(12), 5430–5449 (2013).
[72]A. A. Kognole, A. H. Aytenfisu, and A. D. MacKerell, J. Mol. Model. **26**(6), 152 (2020).
[73]H. Goel, W. B. Yu, V. D. Ustach, A. H. Aytenfisu, D. L. Sun, and A. D. MacKerell, Phys. Chem. Chem. Phys. **22**(13), 6848–6860 (2020).
[74]H. Zheng, M. Chruszcz, P. Lasota, L. Lebioda, and W. Minor, J. Inorg. Biochem. **102**(9), 1765–1776 (2008).
[75]J. A. Lemkul, in *Progress in Molecular Biology and Translational Science*, edited by B. Strodel and B. Barz (Academic Press, 2020), Vol. 170, pp. 1–71.
[76]T. Dudev, Y. L. Lin, M. Dudev, and C. Lim, J. Am. Chem. Soc. **125**(10), 3168–3180 (2003).
[77]E. Pidcock and G. R. Moore, J. Biol. Inorg Chem. **6**(5-6), 479–489 (2001).
[78]J. Gsponer, J. Christodoulou, A. Cavalli, J. M. Bui, B. Richter, C. M. Dobson, and M. Vendruscolo, Structure **16**(5), 736–746 (2008).
[79]C. M. Shepherd and H. J. Vogel, Biophys. J. **87**(2), 780–791 (2004).
[80]O. Y. Hui and H. J. Vogel, Biometals **11**(3), 213–222 (1998).
[81] R. W. Wheatley, D. H. Juers, B. B. Lev, R. E. Huber, and S. Y. Noskov, Phys. Chem. Chem. Phys. **17**(16), 10899–10909 (2015).
[82]C. S. Babu and C. Lim, J. Phys. Chem. A **110**(2), 691–699 (2006).
[83]H. Nguyen, D. A. Case, and A. S. Rose, Bioinformatics **34**(7), 1241–1242 (2018).

Supplementary Figures and Tables for

Benchmarking polarizable and nonpolarizable force fields for Ca^{2+}-peptides against a comprehensive QM dataset

Kazi S. Amin, Xiaojuan Hu, Dennis R. Salahub, Carsten Baldauf, Carmay Lim and Sergei Noskov

Table SI 1: The most common nearest atom-types per residue. To obtain the frequency column for a given dipeptide, we first determined the 3 closest atom-types to Ca^{2+} for each conformation. Then we made a frequency chart of all possible pairs of atom-types from this data. This table reports the most frequent pair for each dipeptide, and its corresponding frequency (number of conformations in which the pair is among the 3 closest atom-types to Ca^{2+} divided by the total number of conformations).

Residue	**Type 1**	**Type 2**	**Frequency (%)**
Ala	OY	CY	66.7
Arg	OY	CY	51.1
Asn	OD1	OY	50.0
Asp	OD2	OD1	87.0
Cys	OY	O	84.1
Gln	OY	OE1	57.4
Glu	OE1	OE2	97.9
HSD	OY	O	56.3
HSE	OY	O	60.4
HSP	OY	CY	74.7
Ile	OY	O	82.1
Leu	OY	O	72.9
Lys	OY	CY	59.1
Met	OY	O	92.6
Phe	OY	O	82.8
Pro	O	OY	66.7
Ser	OY	O	50.0
Thr	OY	O	51.8
Trp	OY	O	70.7
Tyr	OY	O	60.7
Val	OY	O	81.0

Table SI 2: **Summary of the parameters used for CTPOL/AMBER10 model for Glu:Ca2+ system.**

Type	**Element**	**α (Å^3)**	a_L**(e/Å)**	b_L**(e)**
51	N	0.922	-0.1240	0.5244
58	O	0.740	-0.2960	1.2282
59	O	0.739	-1.2054	5.0010
61	O	0.731	-0.3002	1.2456
118	N	0.920	-0.1778	0.7516
127	O	0.730	-0.8433	3.4988
128	O	0.730	-0.6495	2.6947
130	O	0.737	-0.1688	0.7004
719	N	0.899	-0.2056	0.8694
726	O	0.735	-0.1952	0.8096
53	C	1.469		
55	C	1.417		
57	C	1.499		
60	C	1.471		
120	C	1.473		
122	C	1.426		
124	C	1.405		
126	C	1.475		
129	C	1.473		
721	C	1.308		
724	C	1.334		
725	C	1.454		
1979	Ca	10.864		

Table SI 3: **Optimized LJ parameters for CTPOL/AMBER10 model for Glu:Ca2+ system.**

Atom Type-Ion Type	σ (nm)	ε (kJ/mol)
118, 1979	0.3161	0
119, 1979	0.2939	0
120, 1979	0.3265	0
121, 1979	0.296	0
122, 1979	0.3249	0
123, 1979	0.2936	0
124, 1979	0.3241	0
125, 1979	0.2923	0
126, 1979	0.3265	0.0957
127, 1979	0.3133	0
128, 1979	0.3133	0.1798
129, 1979	0.3267	0
130, 1979	0.3142	0.0914
719, 1979	0.3151	0
720, 1979	0.2914	0
721, 1979	0.3206	0
722, 1979	0.2894	0
723, 1979	0.2904	0.2402
724, 1979	0.3217	0
725, 1979	0.326	0
726, 1979	0.3142	0.1202

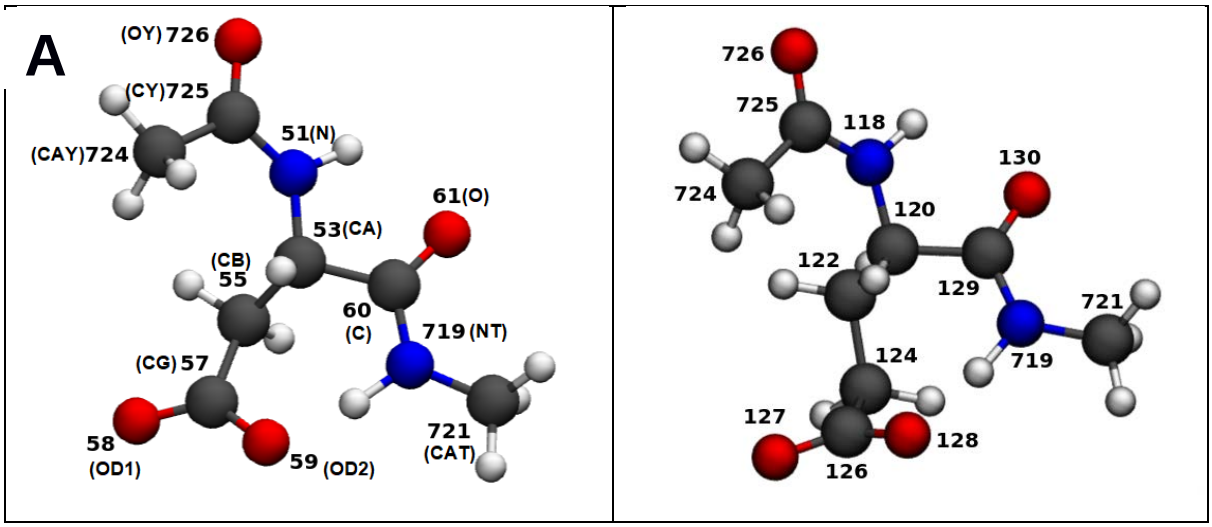


FIG. SI 1: Definition of atom types used in CTPOL (numbers) and C36/Drude (parentheses) for A) Asp and B) Glu.

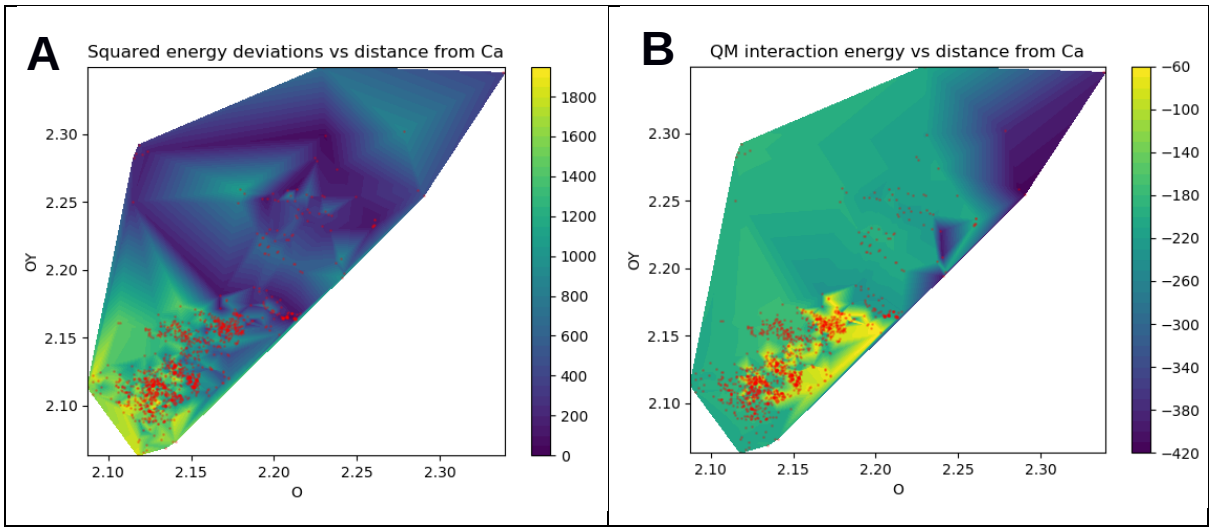


FIG. SI 2: (A) Drude FF squared energy deviations vs distance (Å) between Calcium and the two backbone carbonyl oxygens (O and OY). The red-dots represent all of the conformations of the 20 dipeptides:Ca^{2+}. The conformations of GLU and ASP that result in polarization catastrophe have been removed in order to reveal the trends for other conformations. (B) The corresponding QM interaction energies projected. Both surfaces are obtained by triangle-based linear interpolation of the data.

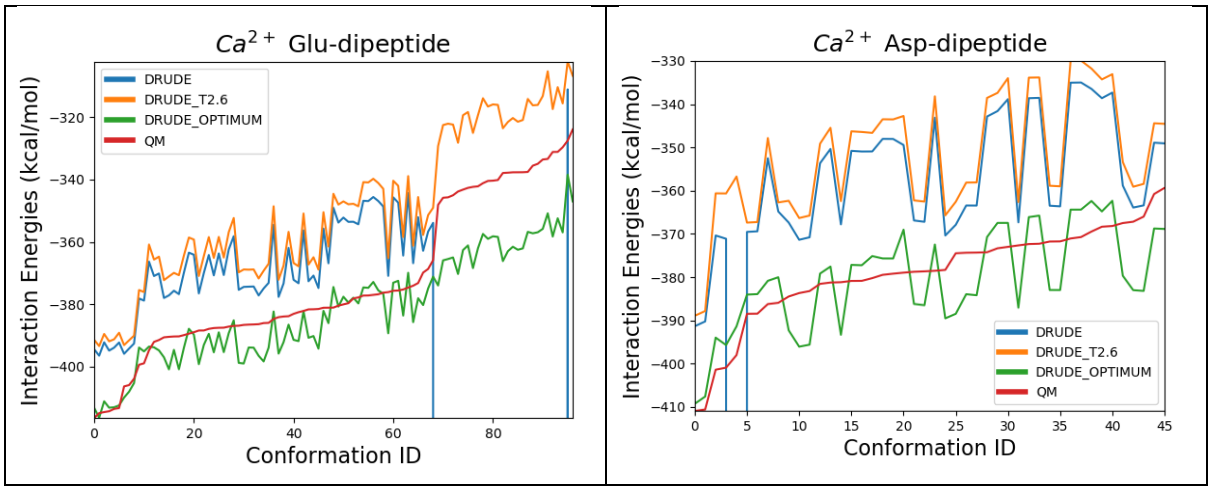
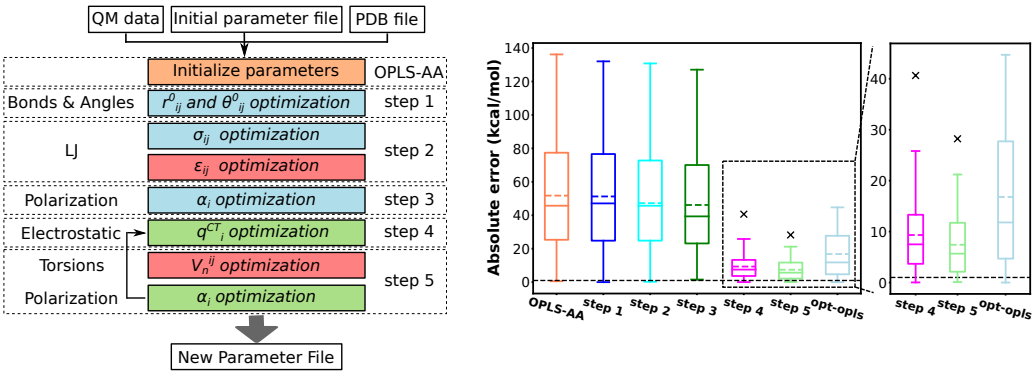


FIG. SI 3: Comparing Drude-FF ion-dipeptide interaction energies before and after parametrization.

4.3 Paper III: System-specific parameter optimization for non-polarizable and polarizable force fields

X Hu, K. S. Amin, M. Schneider, C. Lim, D. Salahub and C. Baldauf.

arXiv preprint arXiv:2303.12775 (2023)



Author contributions: M. Schneider and I developed the parameterization tool FFAFFURR. I did the parameterization of OPLS-AA and CTPOL models and ran the MD simulations of zinc finger protein with different parameter sets. K.S. Amin and I performed the analysis of the MD simulations. C. Baldauf designed the study. C. Lim helped come up with the idea for this study. K.S. Amin, C. Baldauf and I wrote the manuscript.

System-specific parameter optimization for non-polarizable and polarizable force fields

Xiaojuan Hu,[*,†] Kazi S. Amin,[*,‡] Markus Schneider,[†] Carmay Lim,[¶,§] Dennis Salahub,[*,‖] and Carsten Baldauf[*,†]

†*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany*

‡*Centre for Molecular Simulation and Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada*

¶*Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan*

§*Department of Chemistry, National Tsing Hua University, Hsinchu 300, Taiwan*

‖*Centre for Molecular Simulation and Department of Chemistry, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada*

E-mail: xhu@fhi-berlin.mpg.de; kazi.amin@ucalgary.ca; dsalahub@ucalgary.ca; baldauf@fhi-berlin.mpg.de

We dedicate this manuscript to Sergei Noskov, who initiated this work and whose much too early death shook us all.

Abstract

The accuracy of classical force fields (FFs) has been shown to be limited for the simulation of cation-protein systems despite their importance in understanding the processes of life. Improvements can result from optimizing the parameters of classical FFs or by extending the FF formulation by terms describing charge transfer and polarization effects. In this work, we introduce our implementation of the CTPOL model in OpenMM, which extends the classical additive FF formula by adding charge transfer (CT) and polarization (POL). Furthermore, we present an open-source parameterization tool, called FFAFFURR that enables the (system specific) parameterization of OPLS-AA and CTPOL models. The performance of our workflow was evaluated by its ability to reproduce quantum chemistry energies and by molecular dynamics simulations of a Zinc finger protein.

Contents

1 Introduction

Metal ions are essential in biological systems and are involved in physiological functions ranging from maintaining their structural stability to directly participating in catalytic activities.[1] Approximately one-third of all proteins contain metal ions.[2] As an abundant cation in the human body,[3] Zinc is known to play an important role in enzyme catalysis or protein folding/stability. In aqueous solutions, Zn^{2+} normally coordinates with six water molecules in an octahedral coordination geometry. However, in a protein environment, Zn^{2+} is often observed to form a tetrahedral coordination structure with four ligating amino acid residues,[4] commonly His and Cys. Due to the nature of electrostatic interactions, Zn^{2+} also tends to be close to negatively charged residues such as Asp or Glu. Zn^{2+} is involved in various biological functions by interacting with these residues. For example, metallothioneins (MTs)[5,6] are present in all living organisms and are involved in various diseases.[7–9] Under physiological conditions, the four mammalian MT isoforms have Zn_3Cys_9 clusters and Zn_4Cys_{11} clusters in their centers as functional groups. Zinc finger proteins are another well-studied class of Zinc-containing proteins. They play essential roles in DNA recognition, regulation of apoptosis, and protein folding.[10,11] The most well characterized Zinc finger proteins feature a binding domain with two Cys and two His residues. The study of the classical Cys_2His_2 Zinc finger structures is crucial for a better understanding of their broader functions.

Molecular dynamics (MD) simulations employing molecular mechanics (MM) are widely used in the study of complex biological processes, such as protein folding, protein dynamics, and enzyme catalysis because of their ability to model systems at atomic scales ranging in sizes from thousands to millions of atoms and time scales of milli-seconds.[12–14] The majority of current MD studies employ classical force fields (FFs) such as OPLS-AA,[15] AMBER,[16] CHARMM[17] and GROMOS.[18] It is a challenge for classical force field models to describe metal–protein interactions due to the strong local electrostatic field and induction effect,[19–24] for example, computer simulation of Zinc-containing proteins has been a long-standing challenge that appears hard to tackle without explicit treatment of charge-transfer or polar-

ization. One approach to improve the accuracy of force fields is to refine the parameters by fitting the model to more and more accurate experimental data or quantum mechanical (QM) calculations. For example, force-matching algorithms[25] were used to fit parameters to reproduce *ab initio* forces. Empirical Continuum Correction (ECC)[26–28] force fields scale the charges to implicitly take electronic polarization into account. Several works[29,30] tune the Lennard-Jones (LJ) parameters or use a 12-6-4 LJ-type model to simulate charge-induced dipole interactions. These efforts have been successful to some extent, however, reparameterization is often time-consuming and labor-intensive. There are a few automatic parameterization tools, for example, CHARMM General Force Field (CGenFF),[31] LigParGen,[32] and Antechamber.[33,34] These programs typically generate missing parameters for a given system based on analogies with atom types and the relevant parameters available in the corresponding FF or through parameter estimation algorithms.[35] However, the accuracy of assigning approximate parameters to a specific system is limited, and parameters already present in a given FF may also need to be optimized. FFparam[36] and ForceBalance[37] enable the tuning of existing FF parameters. All these parameterization tools share a common assumption of transferability, which assumes a set of parameters optimal for small organic molecules for a given atom type can be applied in a wide range of chemical and spatial contexts. It is well known that the presence of electron donors and acceptors can significantly affect molecular properties by polarization effects.[38] LJ parameters are also sensitive to the local environment[39,40] and long-range electrodynamic screening.[41] In this regard, a fundamentally different approach to derive environment-specific or molecule-specific parameters is proposed in references.[42–44] However, parameters still remain fixed despite structures and environments changing over the course of, e.g., MD simulations.

Another approach to improve FF accuracy in metalloprotein simulations is to introduce more physics to the model. Including polarization effects is a significant step to improve force fields.[45,46] There is growing evidence that polarizable force fields describe ionic systems more accurately than classical force fields. It has been found that the inclusion of polarization

plays an important role in the simulation of ion channels,[47] enzyme catalysis,[48] protein-ligand binding affinity[49] and dynamic properties of proteins.[50]

At present, there are three main groups of polarizable force fields, fluctuating charge, induced point dipoles, and Drude oscillator models.[51] The fluctuating charge models simulate polarization effects by allowing charge to flow through the molecule until the electronegativities of atoms become equalized, while keeping the total charge unchanged.[52] One drawback of the fluctuating charge model is that it fails to capture out-of-plane polarization of planar or linear chemical groups. The fluctuating charge formula can also be used in conjunction with induced point dipoles as a complementary approach to account for charge transfer (CT).[53] A notable model is SIBFA (Sum of Interactions Between Fragments *Ab initio* Computed).[54]

The induced point dipole models describe polarization energy as the interaction between static point charges and induced dipole moments. Notable induced point dipole models include OPLS/PFF,[55] AMBER ff02,[56] and AMOEBA.[57,58] The performance of the induced point dipole models strongly depends on the accuracy of polarizability parameters.

The Drude oscillator model simulates the distortion of the electron density by attaching additional charged particles (the oscillators) to each polarizable atom. Despite many successes of the Drude oscillator model,[19,59,60] it may be limited when charge transfer between cation and coordinating ligand atoms is significant, for example, Cys^- coordinated to metal ions.[61] Ngo *et al.*[62] and Dudev *et al.*[63] showed that the charge located on the coordinating ligand is significantly perturbed due to the presence of Ca^{2+}. The effect exists not only in the first coordination shell, but also in the second shell. Thus, including the description of charge transfer is critical for the development of next-generation polarizable FFs.

The CTPOL[64,65] model incorporates charge transfer (CT) and polarization effects (POL) into classical force fields. The inclusion of charge transfer reduces the amount of partial charge on cation and cation coordinating atoms. Thus, their charge/dipole–charge interactions are weakened. Local polarization energy between cation and coordinating ligands, which also depends on the partial charge, is introduced for compensation.

Although numerous studies have shown that polarizable models perform better than classical force fields in the simulation of metalloproteins, they have received only limited validation. Therefore, reparameterization may be necessary when applied to different systems. Our previous study[21] has shown how QM data[66,67] drive the parameter development of Drude and CTPOL models. However, most parameterization tools focus on classical force field models. FFparam[36] provides parameterization of Drude model; a CTPOL parameterization tool is not yet available.

In this work, we implemented the CTPOL model in OpenMM.[68] The code of this implementation is shared on github.[69] Furthermore, we present a new open-source tool, FFAFFURR (Framework For Adjusting Force Fields Using Regularized Regression), which facilitates the parameterization of OPLS-AA and CTPOL models for a specific system in question, e.g. peptide system or peptide-cation system. One advantage of FFAFFURR is the rapid construction of FFs for troublesome metal centers in metalloproteins. In this work, the new parameters obtained from FFAFFURR are validated by the comparison of FF energies and QM potential energies and MD simulations in the condensed phase using a Zinc finger protein as an example.

2 Methods

2.1 OPLS-AA functional form

OPLS-AA is one of the major families of classical force fields. It is used as the starting point of parameterization in this work. OPLS-AA uses the harmonic functional form to represent the potential energy shown in eq. 1.

$$E^{\mathrm{FF}} = E_{\mathrm{bonds}} + E_{\mathrm{angles}} + E_{\mathrm{torsions}} + E_{\mathrm{improper}} + E_{\mathrm{vdW}} + E_{\mathrm{ele}} \tag{1}$$

where E^{FF} is the potential energy of the system. E_{bonds}, E_{angles}, E_{torsions} and E_{improper} correspond to bonded or so-called covalent terms of bond stretching, bond-angle bending, dihedral-angle torsion, and improper dihedral-angle bending (or out-of-plane distortions) in the molecules. E_{vdW} and E_{ele} are nonbonded terms. They describe van der Waals (vdW) and Coulomb (electrostatic) interactions, respectively.

The energy terms in eq. 1 are depicted in detail in eq. 2.

$$E^{\mathrm{FF}} = \sum_{\mathrm{bonds}}^{1-2\mathrm{atoms}} \frac{1}{2} K_{ij}^{r} \left(r_{ij} - r_{ij}^{0}\right)^{2} + \sum_{\mathrm{angles}}^{1-3\mathrm{atoms}} \frac{1}{2} K_{ij}^{\theta} \left(\theta_{ij} - \theta_{ij}^{0}\right)^{2} + \sum_{\mathrm{dihedrals,n}}^{1-4\mathrm{atoms}} V_{n}^{ij} \left(1 + cos\left(n\phi_{ij} - \phi_{ij}^{0}\right)\right) + \sum_{\mathrm{improper}}^{1-4\mathrm{atoms}} V_{2imp}^{ij} \left(1 + cos\left(2\phi_{ij} - \phi_{ij}^{0}\right)\right) + \sum_{i<j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] f_{ij} + \sum_{i<j} \frac{q_i q_j}{r_{ij}} f_{ij} \tag{2}$$

where K_{ij}^{r}, K_{ij}^{θ}, V_{n}^{ij}, and V_{2imp}^{ij} are force constants, r_{ij}^{0} and θ_{ij}^{0} are the reference bond length and bond angle, r_{ij}, θ_{ij} and ϕ_{ij} are current bond length, bond angle and dihedral angle, respectively, n is the periodicity, ϕ_{ij}^{0} is the phase offset, σ_{ij} is the distance at zero energy, ε_{ij} sets the strength of the interaction, q_i and q_j are the charges of the two particles, and f_{ij} is the scaling factor for short distances (i.e. "1-4 pairs") of nonbonded interaction. In OPLS-AA, the pairwise LJ parameters σ_{ij} and ε_{ij} are calculated as the geometric mean of those of individual atom types (σ_i and ε_i).

Classical force field simulations were performed using OpenMM7, a high performance toolkit for molecular simulations.[68]

2.2 CTPOL model

The CTPOL[64,65] model introduces charge transfer and polarization effects into classical force fields. Instead of a fixed-charge model, CTPOL model takes the charge transfer from ligand atoms L (O, S, N) to metal cation into account. The amount of transferred charge, $\Delta q_{\mathrm{L-Me}}$,

is assumed to depend linearly on the inter-atomic distance, $r_{\mathrm{Me-L}}$

$$\Delta q_{\mathrm{L-Me}} = a_L r_{\mathrm{Me-L}} + b_L. \tag{3}$$

The charge transfer is negligible at distances greater than the sum of the vdW radii of atoms i and j, r_{ij}^{vdW}. Thus, charge on ligand atom L, q_L, can be calculated as

$$q_L = q_L^0 + \Delta q_{\mathrm{L-Me}}, \tag{4}$$

where q_L^0 refers to the charge on atom L in a fixed-charge model.

Polarization energy, E_r^{pol}, can be computed as

$$E_r^{\mathrm{pol}} = -\frac{1}{2}\sum_i \boldsymbol{\mu_i} \cdot \boldsymbol{E_i^0}, \tag{5}$$

where $\boldsymbol{\mu_i}$ is the induced dipole on atom i and $\boldsymbol{E_i^0}$ is the electrostatic field produced by the current charge distribution in the system at the polarizable site i. The summation is over the metal and the metal-bonded residues. A cutoff distance r^{cutoff}, which is equal to the sum of the vdW radii of atoms i and j scaled by a parameter $\gamma = 0.92$, is introduced to avoid unphysically high induced dipoles at close distance. If the distance between atom i and j, r^{ij}, is smaller than r^{cutoff}, we set r^{ij} equal to r^{cutoff}. The only parameter here is the atomic polarizability:

$$\boldsymbol{\mu_i} = \alpha_i \boldsymbol{E_i}, \tag{6}$$

where E_i is the total electrostatic field on atom i due to the charges and induced dipoles in the system.

In this work, we have implemented the CTPOL model on OpenMM via a python script, which can be found at https://github.com/XiaojuanHu/CTPOL_MD.[69] This represents a proof-of-concept implementation, which runs on CPUs. Further code optimization and a transfer to GPUs will likely speed up simulations substantially.

2.3 Reference data set

To evaluate the performance of the parameterization protocol on dipeptide and dipeptide-cation systems, we created a quantum chemistry data set. The data set consists of six models: (1) $AcAla_2NMe$; (2) $AcAla_2NMe{+}Na^+$; (3) deprotonated cysteine: $AcCys^-NMe$, which often plays as the interaction center of metalloproteins; (4) $AcCys^-NMe{+}Zn^{2+}$; (5) $AcCys_2^-NMe{+}Zn^{2+}$, and (6) $AcHisDNMe{+}Zn^{2+}$. The structures and energy hierarchies are shown in Figure 1. The data set can be found on the NOMAD repository via the DOI: 10.17172/NOMAD/2023.02.03-1.[70]

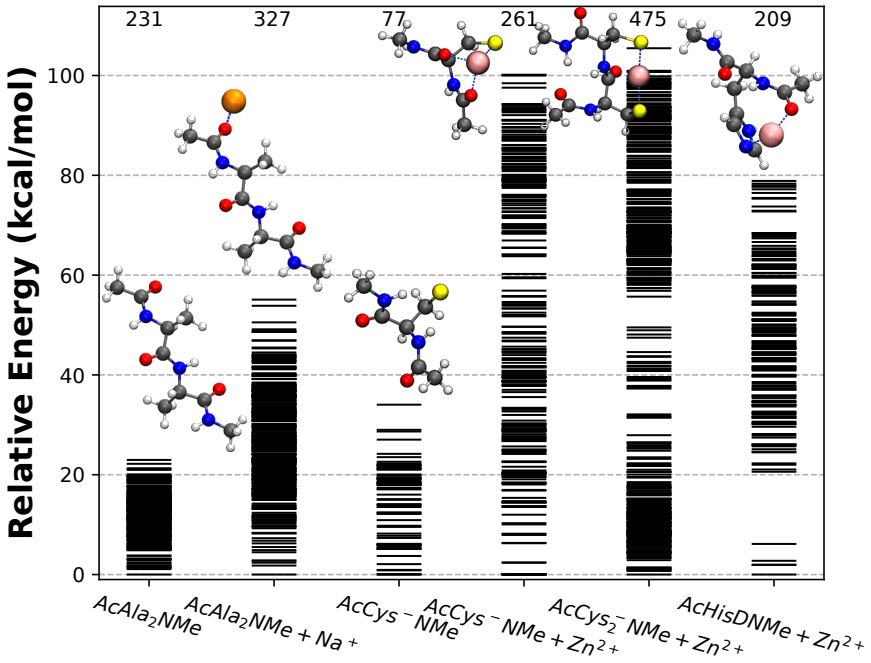


Figure 1: Structures and energy hierarchies of reference data in this study.

All DFT calculations in this work were performed with the numerical atom-centered basis set all-electron code FHI-aims.[71–73] The PBE[74] generalized-gradient exchange-correlation functional augmented by the correction of van der Waals interactions using the Tkatchenko-Scheffler formalism[75] (PBE+vdW^{TS}) was employed. The choice of functional has been validated in previous articles.[66,76] For each conformation, several types of partial charges were provided. Hirshfeld charges[77] are derived based on the Hirshfeld partitioning scheme.[77,78]

ESP charges[77,79] are derived by fitting partial charges to reproduce the electrostatic potential. RESP charges[80] are extracted by a two-stage restrained electrostatic potential (RESP) fitting procedure[80] within the Antechamber suite of the AmberTools package.[16] The electrostatic potential was evaluated on a set of grids in a fixed spatial region located in a cubic space around the molecule. The 5 radial-shells were generated in a radial region between 1.4 and 2.0 multiples of the atomic vdW-radius. The cubic space contains 35 points along x, y, and z directions, respectively.

The conformers of $AcAla_2NMe$, $AcAla_2NMe{+}Na^+$, and $AcHisDNMe{+}Zn^{2+}$ were obtained by a conformational search algorithm as shown in the studies of Rossi *et al.*[81] and Schneider *et al.*[23] First, a global conformational search was performed with the basin-hopping approach[82,83] at the force field level (OPLS-AA).[84] The scan program of the TINKER molecular modeling package[85,86] was employed to perform the basin-hopping search strategy. An energy threshold of 100 kcal/mol for local minima and a convergence criterion for local geometry optimizations of 0.0001 kcal/mol were used. All obtained conformers were relaxed at PBE+vdW^{TS} level with *tier 1* basis set and *light* setting employed. A clustering scheme was then applied to exclude duplicates using the root-mean-square deviations (RMSD) of atomic positions. Finally, further relaxation was accomplished at the PBE+vdW^{TS} level using *tier 2* basis set and *tight* setting.

The conformers of $AcCys^{-}NMe$, $AcCys^{-}NMe{+}Zn^{2+}$, and $AcCys_2^{-}NMe{+}Zn^{2+}$ were obtained with the genetic algorithm (GA) package Fafoom.[87] First, a GA search at the PBE+vdW^{TS} level with *light* basis set was employed for structure sampling. Then a clustering scheme with a clustering criterion of RMSD of 0.02 Å for atomic positions and a relative energy of 0.02 kcal/mol was applied to remove duplicates. The obtained conformers were further relaxed with FHI-aims[71–73] at the PBE+vdW^{TS} level with *tight* basis set. Final conformers were obtained after clustering. Both conformational search protocols have been well validated.[81,87]

2.4 Parameter optimization

Optimization methods used in this work include LASSO (least absolute shrinkage and selection operator)[88] regression, Ridge regression[89] and particle swarm optimization (PSO).[90,91] If the parameters enter the force field function in a quadratic way, e.g. V_n^{ij}, the optimization can be performed by solving a set of linear equations. In this case, LASSO and Ridge regression were employed to treat the potential overfitting. The regularization parameter λ in LASSO and Ridge regression was selected by 10-fold cross-validation. LASSO and Ridge regression were performed with Python's scikit-learn[92] library. If the parameters can not be obtained by solving a set of linear equations, e.g. charge transfer parameters a_L, PSO was employed. Similar to GA, PSO is a powerful population-based global optimization algorithm. It relies on a population of candidate solutions, called particles, and finds the optimal solution by moving these particles through a high-dimensional parameter space based on their position and velocity. PSO was performed with the python package pyswarm.[93]

2.5 FFAFFURR

Force field parameterization in principle has three iterative and challenging steps:[94]

1) Definition of the optimization problem (selection of reference data, objective of the optimization, and force field parameters to adjust): High quality QM data has been used for FF parameterization and is likely continue to be an essential part of next-generation FF development.[95] FFAFFURR uses high quality QM data as described in section 2.3 as the reference. In principle, the parameters of every energy term in a force field have to be optimized since the parameters of all terms are interdependent, only adjusting one energy term may cause parameter inconsistency. Users can choose which energy terms to tune according to specific problems. OPLS-AA parameters are used as initial parameters.

2) Force field parameterization: The framework and algorithms used in FFAFFURR are

explained in this section.

3) Validation of optimized parameters: The performance of the FF parameter sets obtained from FFAFFURR is evaluated by the ability to reproduce the DFT (or any other high-level method) potential energies and by the MD simulations.

Some practical points were considered when establishing the FFAFFURR framework: (i) the framework should be straightforward to set up and use, (ii) it should be easy to extend with other FF parameters or functional forms, and (iii) the result should be immediately usable by a molecular simulation package. FFAFFURR acts as a "wrapper" between the molecular mechanics package openMM[68] and the *ab initio* molecular simulation package FHI-aims.[71–73] The code reads QM data directly from the output of FHI-aims and the output itself is a parameter file that can be processed by openMM. FFAFFURR is designed as the next step of the genetic algorithm package Fafoom.[87] Conformers obtained by Fafoom through global search can be directly parsed to FFAFFURR. FFAFFURR is an open source tool and can be found at https://github.com/XiaojuanHu/ffaffurr-dev/releases/tag/version1.0.

2.5.1 Bond and angle parameterization

K_{ij}^{r}, K_{ij}^{θ}, r_{ij}^{0} and θ_{ij}^{0} are empirical parameters of bond-stretching and angle-bending terms. The "spring" parameters K_{ij}^{r} and K_{ij}^{θ} are unaltered in FFAFFURR. The focus simply lies on the "torsional" and "non-bonded" parameters. Bond-stretching and angle-bending terms intend to model small displacements away from the lowest energy structure. We adjust r_{ij}^{0} and θ_{ij}^{0} by simply taking the average of the respective bond or angle over all local minima in the quantum chemistry data set.

2.5.2 Torsion angle parameterization

The torsion angle term represents a combination of the bonded and nonbonded interactions. It has been reported that torsional parameters fitted to gas phase QM data perform similarly

to those fitted to the experimental data.[95] Although torsional parameters can be derived from vibrational analysis or using vibrational spectra as target data, this approach is complicated and requires a more elaborate treatment.[36,96,97] In the case of the torsion term, force constants V_n^{ij} and V_{2imp}^{ij} can be tuned by LASSO or Ridge regression to minimize the difference between the FF and QM torsional energies. The "torsions contribution" from QM $\widetilde{E}_{\text{torsions}}^{\text{QM}}$ is calculated as:

$$\widetilde{E}_{\text{torsions}}^{\text{QM}} = E_{\text{total}}^{\text{QM}} - E_{\text{nonbonded}}^{\text{FF}} - E_{\text{bond}}^{\text{FF}} - E_{\text{angle}}^{\text{FF}}, \tag{7}$$

where $E_{\text{total}}^{\text{QM}}$ represents the total energy of conformer from QM calculation, $E_{\text{nonbonded}}^{\text{FF}}$, $E_{\text{bond}}^{\text{FF}}$ and $E_{\text{angle}}^{\text{FF}}$ represent energies of nonbonded terms, bond term, and angle term from FF calculation, respectively.

2.5.3 Electrostatic parameterization

A key difference between FFs is how they derive atomic partial charges. Deriving charges from QM data is widely used. The workflow of FFAFFURR tested three choices of partial charges: Hirshfeld,[77,78] ESP[77,79] and RESP[80] charges. The charge of each atom type of the force field is defined as the average value of QM charges. The scaling factor f_{ij} used to scale the electrostatic interactions between the third neighbors (1,4-interactions) can also be adjusted by fitting to minimize the difference between the FF and QM energies.

2.5.4 LJ parameterization

Pair-specific Lennard−Jones (LJ) interaction parameters (referred to as NBFIX in the CHARMM force fields) have been proven to better describe the interaction between cations and carbonyl groups of a protein backbone.[19] FFAFFURR employs pairwise Lennard−Jones (LJ) parameters instead of values determined by the combination rule.

In recent years, progress has been made in the calculation of pairwise dispersion interaction strength from the ground-state electron density of molecules.[98–100] The interatomic pairwise parameter σ_{ij} can be derived using the atomic Hirshfeld partitioning scheme, which

has already been used in the pairwise Tkatchenko-Scheffler vdW model. With the concept of the vdW radius, the LJ energy can be written as

$$E_{\text{vdw}} = \sum_{i<j} \varepsilon_{ij} \left[\left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{6} \right] f_{ij}, \tag{8}$$

where R_{ij}^{min} refers to the atomic distance where the vdW potential is at its minimum. With the definition of the effective atomic volume, R_{ij}^{min} is estimated as the sum of effective atomic van der Waals radii of atom i and atom j. The effective vdW radius of an atom is given by

$$R_{\text{eff}}^{0} = \left(\frac{V^{\text{eff}}}{V^{\text{free}}} \right)^{1/3} R_{\text{free}}^{0}, \tag{9}$$

where R_{free}^{0} is the free-atom vdW radii that correspond to the electron density contour value determined for the noble gas on the same period using its vdW radius by Bondi.[101] Pairwise σ_{ij} can be calculated as

$$\sigma_{ij} = 2^{-1/6} R_{ij}^{\text{min}}. \tag{10}$$

The ε_{ij} parameter from eq. 8 can be tuned by fitting FF LJ energies to reproduce QM vdW energies by LASSO or Ridge regression.

2.5.5 Deriving charge transfer parameters

In all Zinc finger proteins and most enzymes, Zn^{2+} coordinates to four ligands. However, due to the setup of the QM data set with monomeric and dimeric peptides, the cations have coordination numbers (CNs) of one or two. Therefore we added a correction factor for CN in eq. 3

$$\Delta q_{\text{L-Me}} = \frac{1}{\text{CN}^{k}} (a_L r_{\text{Me-L}} + b_L). \tag{11}$$

k, a_L, and r^{cutoff} can be adjusted by PSO. The target objective of fitting can be QM potential energy, QM interaction energy, or electrostatic potential. b_L can be calculated with the assumption that charge transfer is zero at cutoff distance.

2.5.6 Polarization energy

To get the value of atomic polarizability α_i in eq 6, we use the definition of effective polarizability of an atom in a molecule, where the free-atom polarizabilitiy is scaled according to its close environment with a partitioning:

$$\alpha_{\text{eff}} = \left(\frac{V^{\text{eff}}}{V^{\text{free}}} \right) \alpha^0_{\text{free}}, \tag{12}$$

where V^{eff} and V^{free} are the same in eq. 9, and α^0_{free} is the isotropic static polarizability. α_i is taken by averaging over all atoms with the same atom type in the quantum chemistry data set. FFAFFURR also supports to slightly adjust α_i by fitting force field energies to reproduce QM energies via PSO.

2.5.7 Boltzmann-type weighted fitting

The quantum chemistry data set covers a wide range of relative energies. By transitioning from, in our case, DFT to an additive force field, even including charge transfer and polarization, we reduce dimensionality of the energy function and therewith to represent the PES. Consequently, a force field, describing, e.g., such a cation-protein system, cannot fully reproduce a DFT PES. Hence, it is advisable to focus on accuracy of certain areas of the PES. RMSD between two surfaces is a common fitting criteria, but this approach gives more weight to areas of the energy surface with larger absolute values, while the real weight should more closely represent the Boltzmann weight of the energy surface. Consequently, we calculate Boltzmann-type weights and apply them as a scoring function. The weighted RMSD, $wRMSD$, is given as:

$$wRMSD = \left[\sum_{i=1}^{N} w_i (E_i^{\text{FF}} - \Delta E_i^{\text{QM}})^2 \right]^{\frac{1}{2}}, \tag{13}$$

where RMSD is modified by including a Boltzmann-type factor,

$$w_i = A \exp\left[\frac{-E_i^{\mathrm{QM}}}{\mathrm{RT}}\right], \tag{14}$$

where A is the normalization constant (so that $\sum w_i = 1$) and RT is the "temperature factor" that has no physical meaning in the context of this application, but affects the flatness of the distribution. Our previous work[21] has shown how Boltzmann-type weighted RMSD with appropriate choice of RT can be utilized as objective function for force field parameter optimization. Therefore, we implemented Boltzmann-type weighted fitting in FFAFFURR by scaling the energies with the corresponding Boltzmann-type weights.

2.6 Validation of new parameters

2.6.1 Assessment of the energies

To evaluate the performance of the parameterization, energies of conformers in the test set calculated with optimized parameters were compared to DFT energies by mean absolute errors (MAEs) and maximum errors (MEs). The MAE for the relative energies between FF energies and QM energies is calculated as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\Delta E_i^{\mathrm{FF}} - \Delta E_i^{\mathrm{QM}} + c|, \tag{15}$$

where N is the number of conformers in a given data set. ΔE_i refers to the energy difference between conformer i and the lowest-energy conformer in the set. The adjustable parameter c is used to shift the FF or QM energy hierarchies to one another to get the lowest MAE. ME is calculated as:

$$\mathrm{ME} = \max_{i \in N} |\Delta E_i^{\mathrm{FF}} - \Delta E_i^{\mathrm{QM}} + c|. \tag{16}$$

2.6.2 Molecular dynamics simulations

We performed MD simulations of the NMR structure 1ZNF[102] with different parameter sets to evaluate the performance of FFAFFURR. All MD simulations were performed using OpenMM7.[68] The structure of 1ZNF was placed in a cubic box of 68 Å side length filled with TIP3P water. Four Cl^- were added to neutralize the system. Then energy minimization was performed with the steepest descent minimization. To equilibrate the solvent and ions around the protein, we continued 100 ps NVT and 100 ps NPT equilibration at a temperature of 300 K. SHAKE constraints were applied to heavy atoms of the protein. Then independent MD simulations were performed with a time step of 2 fs. In all calculations, the long-range electrostatics beyond the cutoff of 12 Å were treated with the Particle Mesh Ewald (PME) method.[103] The LJ cutoff was set to 12 Å. The LJ and electrostatic interactions were computed every time step. For the simulations with the CTPOL model, charge transfer and induced dipoles were updated every 10 steps. Covalent bonds and water angles were constrained.

3 Results and discussion

To assess the performance of FFAFFURR and describe which protocol to use to create the parameter set, we optimized the parameters of OPLS-AA with FFAFFURR and extended the OPLS-AA model by the CTPOL model. The quality of optimized parameters was assessed by assessing the structural stability of the Zinc finger motif in MD simulations.

3.1 OPLS-AA parameterization

Although studies have shown that it is difficult to implicitly incorporate the polarization effect into classical FFs,[21,104] fine-tuning parameters of fixed-charge models to describe cation-protein systems is still attractive due to its low computational cost and easier parameterization. Here we tested the performance of the fixed-charge model OPLS-AA parametrized

by FFAFFURR. Five systems were tested: (1) $AcAla_2NMe$; (2) $AcAla_2NMe+Na^+$; (3) $AcCys^-NMe$; (4) $AcCys^-NMe+Zn^{2+}$; and (5) $AcCys_2^-NMe+Zn^{2+}$. $AcAla_2NMe$ and $AcAla_2NMe + Na^+$ were used as toy models since the polarization effect caused by Na^+ is minor. On the contrary, Cys^- is one of the ligands that interact with Zn^{2+} in proteins, and charge transfer between Cys^- and Zn^{2+} is significant. For each system, 80 percent of the conformers were randomly selected as the training set, and the remaining 20 percent were used as the test set.

We first demonstrate the functionality of FFAFFURR on the example of OPLS-AA parameterization. The key steps of OPLS-AA parameterization are briefly described in Figure 2 (a). We showed the ability to reproduce PES by optimizing parameters of bonds, angles, electrostatic interactions, LJ interactions, and torsional interactions. Users can choose which energy items to adjust according to their needs. In Figure 2 (a), the parameters in blue boxes are derived from DFT calculations and the parameters in coral boxes are fitted by LASSO or Ridge regression as described in Section 2.5. Here, we only tested RESP partial charges, LASSO method in ε_{ij} deriving, and Ridge regression in V_n^{ij} deriving. The parameterization protocol followed the order shown in Figure 2 (a).

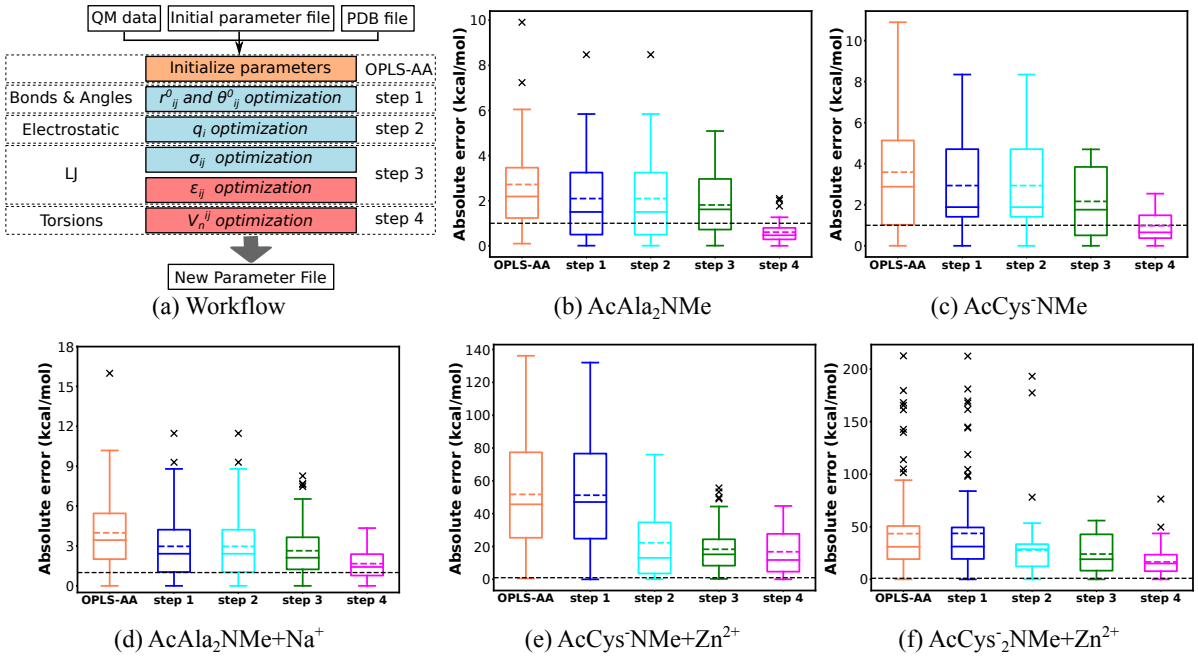
Figure 2 (b-f) shows the comparison of FF energies with optimized parameters after each step in Figure 2 (a) to QM energies. Noticeably, charges for $AcAla_2NMe$, $AcCys^-NMe$ and $AcAla_2NMe+Na^+$ were not altered since the original charges yielded errors lower than average RESP charges from QM calculations, while average RESP charges were employed for $AcCys^-NMe+Zn^{2+}$ and $AcCys_2^-NMe+Zn^{2+}$. Figure 2 (e) and (f) indicate that using average RESP charges significantly reduces absolute errors for $AcCys^-NMe+Zn^{2+}$ and $AcCys_2^-NMe+Zn^{2+}$. This could be due to the capture of charge transfer to some extent. In the case of $AcAla_2NMe$ and $AcCys^-NMe$, the MAEs were improved from 2.72 kcal/mol and 3.59 kcal/mol to 0.61 kcal/mol and 0.98 kcal/mol, respectively, which are better than the chemical accuracy 1 kcal/mol. In the case of $AcAla_2NMe+Na^+$, the MAE was improved from 3.99 kcal/mol to 1.67 kcal/mol. Although the optimized MAE is higher than

the chemical accuracy, the maximum error is significantly reduced. However, in the cases of $AcCys^{-}NMe{+}Zn^{2+}$ and $AcCys_2^{-}NMe{+}Zn^{2+}$, the MAEs were improved from 51.75 kcal/mol and 43.47 kcal/mol to 16.8 kcal/mol and 16.59 kcal/mol, respectively. Although this is a great improvement, the MAEs are much higher than other systems, the calculations based on these parameters still have no predictive power. This confirms the necessity of explicitly including charge transfer and polarization effects to describe the divalent ion-dipeptide systems. We note that for dipeptides and dipeptides with monovalent cation systems, the greatest influence factor is the optimization of torsional parameters. Previous studies by some of us[76,105] have shown that cations strongly modify the preferences of torsion angles. While for dipeptides with divalent cations, the adjusting of charge plays the most important role. This further confirms that the capture of charge transfer and polarization is crucial for the accurate description of systems with divalent cation. We also note that the maximum errors are greatly reduced after the parameterization of LJ interactions of the five systems.

3.2 CTPOL parameterization

The CTPOL model introduces both local polarization and charge-transfer effects into classical force fields. We investigated the performance of the CTPOL model on the cation-dipeptide systems: $AcAla_2NMe{+}Na^{+}$, and two challenging systems $AcCys^{-}NMe{+}Zn^{2+}$ and $AcCys_2^{-}NMe\ {+}Zn^{2+}$. The major steps of the CTPOL parameterization workflow are depicted in Figure 3 (a). As already mentioned, the parameters in blue boxes are derived from DFT calculations and the parameters in coral boxes are fitted by LASSO or Ridge regression. Furthermore, the parameters in green boxes are obtained by PSO. Noticeably, α_i is tuned twice. In step 3, α_i is taken as the average effective polarizability calculated from *ab initio* method. In step 5, we tried to slightly tune α_i by PSO. An additional round of parameterization from step 4 to step 5 can be performed to better optimize the FF parameters.

Absolute errors of each step in Figure 3 (a) are illustrated in Figure 3 (b-f). Absolute errors of optimized OPLS-AA (opt-opls) are also shown in Figure 3 to compare the per-



(a) Workflow (b) $AcAla_2NMe$ (c) $AcCys^-NMe$

(d) $AcAla_2NMe+Na^+$ (e) $AcCys^-NMe+Zn^{2+}$ (f) $AcCys^-_2NMe+Zn^{2+}$

Figure 2: (a) Workflow of the parameterization of OPLS-AA in four major steps. Different colors represent different fitting methods. Parameters in blue boxes are derived from DFT calculation, and parameters in coral boxes are tuned by LASSO or Ridge regression. (b-f) Box plots of absolute errors of OPLS-AA parameterization major steps (OPLS-AA, step 1, step 2, step 3, step 4) for the test set of (b) $AcAla_2NMe$, (c) $AcCys^-NMe$, (d) $AcAla_2NMe+Na^+$, (e) $AcCys^-NMe+Zn^{2+}$ and (f) $AcCys^-_2NMe+Zn^{2+}$. The upper and lower lines of the rectangles mark the 75% and 25% percentiles of the distribution, the horizontal line in the box indicates the median (50 percentile), internal colored dash line indicate the mean value, and the upper and lower lines of the "error bars" depict the 99% and 1% percentiles. The crosses represent the outliers. Black dash line indicates the chemical accuracy, which is 1 kcal/mol.

formance of FFAFFURR on OPLS-AA and CTPOL models. As shown in Figure 3, the introduction of polarization effects in step 3 didn't improve the accuracy much, and the errors of $AcAla_2NMe{+}Na^+$ system even increased. This may be due to the fact that classical force fields already include part of polarization effect, since the charges come from fitting to reproduce quantum mechanical or experimental electrostatic field distribution.[65] Including charge transfer from ligand atoms to cation reduces atomic charges, therefore compensating for the electrostatic potential. Not surprisingly, errors are significantly reduced after including charge transfer as displayed in Figure 3. After the parameterization, the MAEs of $AcAla_2NMe{+}Na^+$, $AcCys^-NMe{+}Zn^{2+}$ and $AcCys_2^-NMe{+}Zn^{2+}$ reached 1.45 kcal/mol, 7.42 kcal/mol, and 8.12 kcal/mol, respectively. In contrast, the MAEs of the optimized OPLS-AA are 1.67 kcal/mol, 16.8 kcal/mol, and 16.59 kcal/mol, respectively. Apparently, the inclusion of charge transfer and polarization effects better describes systems involving cations than classical force fields, especially for systems with divalent cations.

3.3 Weighted fitting

To focus the fitting on the low energy part of the PES, we applied Boltzmann-type weights to the scoring function during the fitting of charge transfer parameters. In Figure 4, $AcCys^-NMe{+}Zn^{2+}$ system is taken as an example. Figure S1 shows the Boltzmann-type weights (w_i) along QM relative energies with different temperature factor (RT) values. The weight decreases as the relative energy increases. And larger RT values put less weight on low energy conformations. Figure 4 shows the difference in mean absolute errors between unweighted fitting and weighted fitting with RT = 16. In Figure 4, the height of the bar represents the mean absolute error for conformers whose relative energies are smaller than the right node of the bar. Interestingly, the weighted fitting improves accuracy a lot in the low-energy region, while high-energy regions do not get worse.

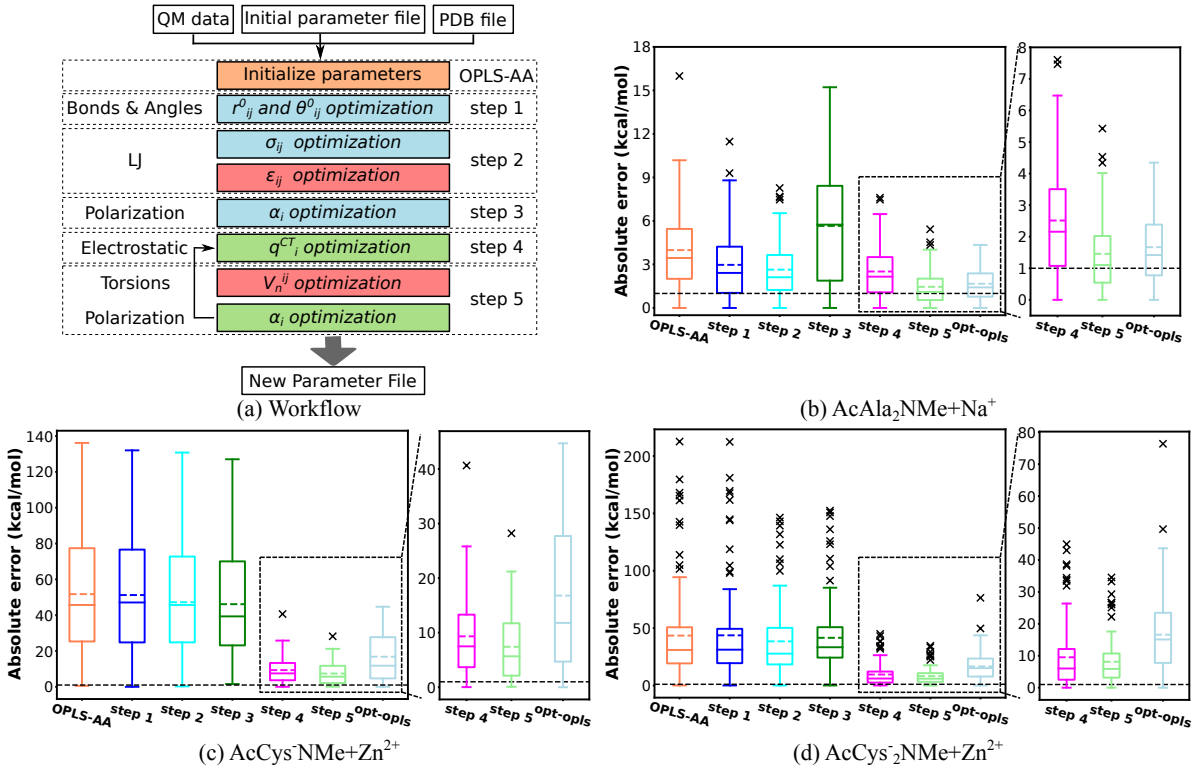


Figure 3: (a) Workflow of full CTPOL parameterization in five major steps. Different colors represent different fitting methods. Parameters in blue boxes are derived from DFT calculation, parameters in coral boxes are tuned by LASSO or Ridge regression, and parameters in green boxes are tuned by PSO. (b-d) Box plots of absolute errors of CTPOL parameterization major steps (OPLS-AA, step 1, step 2, step 3, step 4, step 5) and OPLS-AA with full optimized parameters (opt-opls) for test set of (b) $AcAla_2NMe{+}Na^+$, (c) $AcCys^-NMe{+}Zn^{2+}$ and (d) $AcCys^-_2NMe{+}Zn^{2+}$. The upper and lower lines of the rectangles mark the 75% and 25% percentiles of the distribution, the horizontal line in the box indicates the median (50 percentile), internal colored dash line indicate the mean value, and the upper and lower lines of the "error bars" depict the 99% and 1% percentiles. The crosses represent the outliers. Black dash line indicates the chemical accuracy, which is 1 kcal/mol.

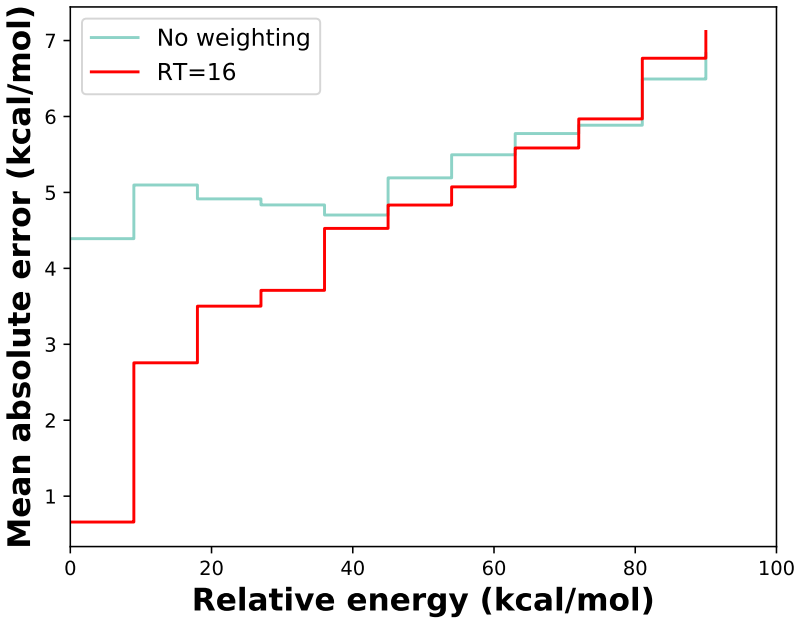
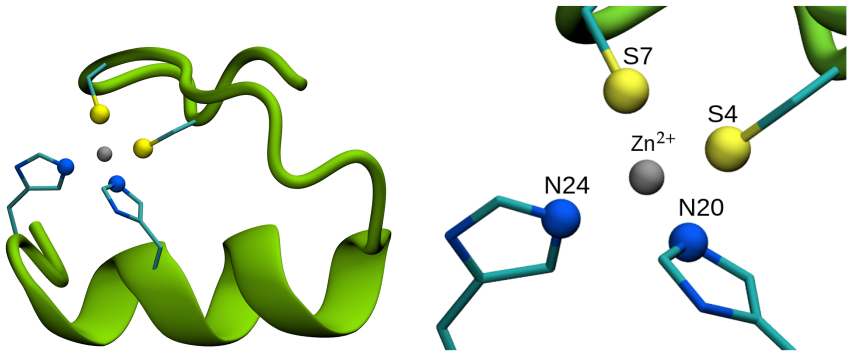


Figure 4: Absolute errors of QM and optimized FF energies by weighted/unweighted fitting of AcCys^{-}NMe+Zn^{2+} system. The height of the bar represents the mean absolute error for conformers whose relative energies are smaller than the right edge of the bar.

3.4 Validation with molecular dynamics simulations

Zinc fingers[106] are extremely common DNA binding motifs found in eukaryotes which coordinate one or more zinc ions.[107] Multiple fingers can combine together to carry out many complex functions, such as regulating DNA/RNA transcription,[106,107] protein folding and assembly, lipid binding, Zinc sensing,[10] and even protein recognition.[108]

The 1ZNF PDB structure[102] is one of the first Zinc finger structures to be resolved experimentally. It is also the simplest, containing only 25 amino acids and one Cys_2His_2 Zinc binding domain where the Zinc ion is in a stable coordination geometry consisting of cysteine sulfurs and histidine nitrogens in the first coordination shell (see Figure 5). Due to its compact size, the 1ZNF structure provides an ideal case study for an MD validation of a FFAFFURR parameterization workflow. One potential application of FFAFFURR to this system is to optimize selected parameters for the interaction center (Figure 5 (b)), since



(a) Protein backbone and Zn^{2+} interaction center.

(b) Close-up of interaction center

Figure 5: A view of the protein structure from the 1st model of the NMR strucutre 1ZNF. The numbers in the atom names refer to the residue number. The sulfurs are from Cys4 and Cys7, while the nitrogens are the NE2 nitrogens of His20 and His24.

that is the region of most complexity.

In this paper, we used an approach similar to Li *et al.*,[109] giving the residues in the interaction center unique residue names to distinguish them from similar residues in the rest of the protein. This allows us to target only atom-types within the binding domain for parameterization, without affecting the parameters of similar atom-types away from the binding site.

Four parameter sets were tested with MD in this study, as described in Table 1. For the unparameterized OPLS-AA force-field, we observed unbinding of the two histidine residues from the Zn^{2+} interaction center after 40 ns of simulation, as shown in Figure 7. To try and prevent this, we parameterized pair-wise LJ parameters between atoms in HisD and Zn^{2+}. The parameters that are optimized are listed in Table S2. The LJ parameters between atoms in Cys and Zn^{2+} are keep untouched since we haven't seen strange behaviors between Cys and Zn^{2+}. The parameterized LJ parameters were used in opt-OPLS-AA and opt-CTPOL sets. In the CTPOL and opt-CTPOL models, charge transfer was introduced for S/N/O atoms in the binding site, and polarization effects between non-hydrogen atoms and Zn^{2+} were added.

Table 1: Parameter sets used for MD simulation. The determination of LJ parameters from FFAFFURR is described in 2.5.4. optimized parameters are listed in Table S2 and S3.

Parameter set	Pair-wise LJ parameters of atoms in HisD and Zn^{2+}	CT + POL
OPLS-AA	original	No
opt-OPLS-AA	from FFAFFURR	No
CTPOL	same as OPLS-AA	Yes
opt-CTPOL	from FFAFFURR	Yes

The protein backbone structure and binding domain are better preserved with charge transfer and polarizability

We ran three 40 ns long simulation with each of the four models listed in Table 1. We also used the 37 experimental NMR structures of 1ZNF to compare structural features between our simulations and NMR observations. Figure 6 shows the RMSD of each of the parameter sets, using the first model of the NMR structures as a reference. In the same figure, we also plot the RMSD of the 37 NMR models with respect to the the same first model to see how much variation occurs among those.

It is clear from Figure 6 that both the overall structure and binding domain are in better agreement with the NMR structures when charge transfer and polarizability are taken into account. With opt-OPLS-AA, there is a marginal but noticeable improvement over OPLS-AA, but in both OPLS-AA and opt-OPLS-AA force fields the binding domain breaks apart. This is evident from the RMSD of the backbone, as shown in the bottom panel of Figure 6. This is primarily due to the Histidines breaking away from the binding with Zn^{2+}, as supported by Figure S2.

The RMSDs of OPLS-AA and opt-OPLS-AA deviate far from the NMR model, particularly the RMSDs of the binding site only. We observed in our simulations that with OPLS-AA, the two histidine residues in the binding site stray uncharacteristically far from Zn^{2+}. Even with optimization the pair-wise LJ parameters of Zn^{2+} and histidine (opt-OPLS-AA), we observed one of the histidines escaping the binding domain. Figure 7 (a) and (b) shows snapshots of such conformations after 40 ns. Similar problems with binding domain stability

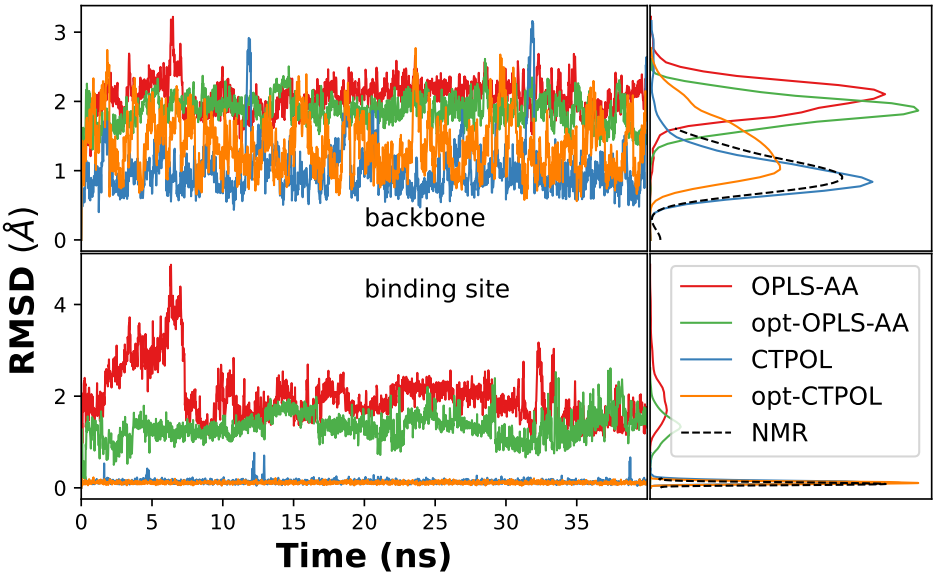
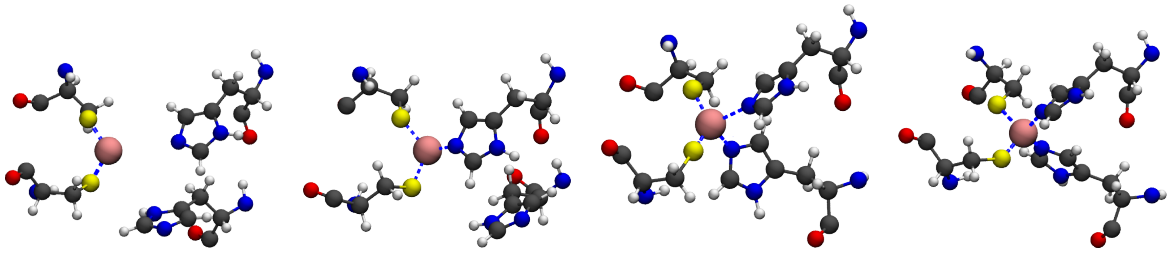


Figure 6: RMSDs of MD trajectories from the NMR structure of 1ZNF (Model 1), calculated for different parameter sets. Top: the protein backbone atoms only. Bottom: the binding site containing Zn, S4, S7, N20, and N24, as shown in Figure 5. The densities of RMSD values are shown on the right, using Kernel Density Approximation,[110,111] where the dashed line is the RMSD distribution obtained from NMR data of 1ZNF with respect to the first model of the PDB.



(a) OPLS-AA (b) opt-OPLS-AA (c) CTPOL (d) opt-CTPOL

Figure 7: Snapshots showing the conformation of binding site after 40 ns of simulation.

have been observed in previous studies, where the Zn^{2+} escapes from the coordination center in non-polarizable FF simulations.[104,112]

However, both CTPOL and opt-CTPOL preserve the binding domain of Zn^{2+}, with both histidines and both cysteines coordinating the Zn^{2+} ion throughout the 40 ns simulations (snapshots of Figure 7 (c) and (d)). This emphasizes that explicitly including charge transfer and polarization effects is critical for a proper description of the binding domain, and hence the overall structure of Zinc fingers.

Lennard Jones parameterization further stabilizes the CTPOL model

To evaluate the effect of optimized pair-wise LJ parameters we compared the CTPOL model without any LJ parameterization (CTPOL) to the CTPOL model with LJ parameterization (opt-CTPOL). From Figure 6, it may appear that such optimization has little effect, and in fact may slightly worsen the overall structure due to the higher RMSD of the backbone. However, while both models preserve the interaction center much better than OPLS-AA and opt-OPLS-AA, opt-CTPOL appears to produce a much more stable binding domain than CTPOL. This can be seen when we recompute RMSD after varying the initial conditions. To test the impact of initial conditions, we ran 40 independent 1ns long simulations, with the initial frame randomly chosen from a 4 ns MD simulation and random initial velocities. These are reasonable initial conditions that should exhibit similar behaviour, as they are taken from a simulation. Figure 8 shows that while the 40 ns trajectory of CTPOL using the crystal structure as starting point is more or less stable, when running simulations from different initial conditions, this stability is not guaranteed, as seen from the spikes in RMSD. On the other hand, opt-CTPOL appears to be stable for all initial conditions.

A reason for this is the abnormal charge transfers to Zinc in CTPOL as seen in Figure 9. This occurs around the same time as the binding domain fluctuations in Figure 8 . A closer inspection of the distances between Zinc and coordinating nitrogens (Figure 10) reveals that these fluctuations are perfectly correlated with these distances. As the binding site breaks

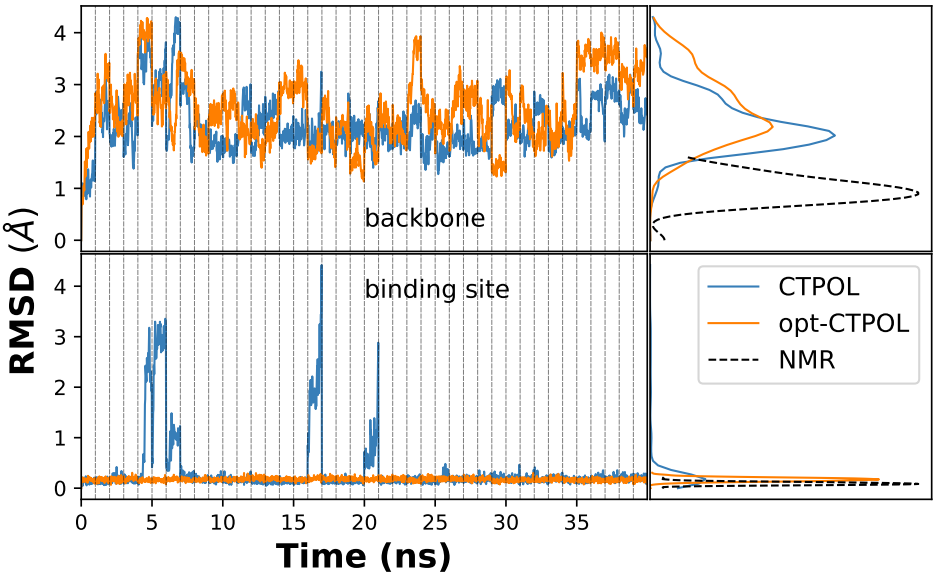


Figure 8: RMSD of CTPOL and opt-CTPOL vs 1st model of NMR, with 40 trajectories of 1 ns concatenated into one. The dotted lines represent concatenation boundaries of the trajectories.

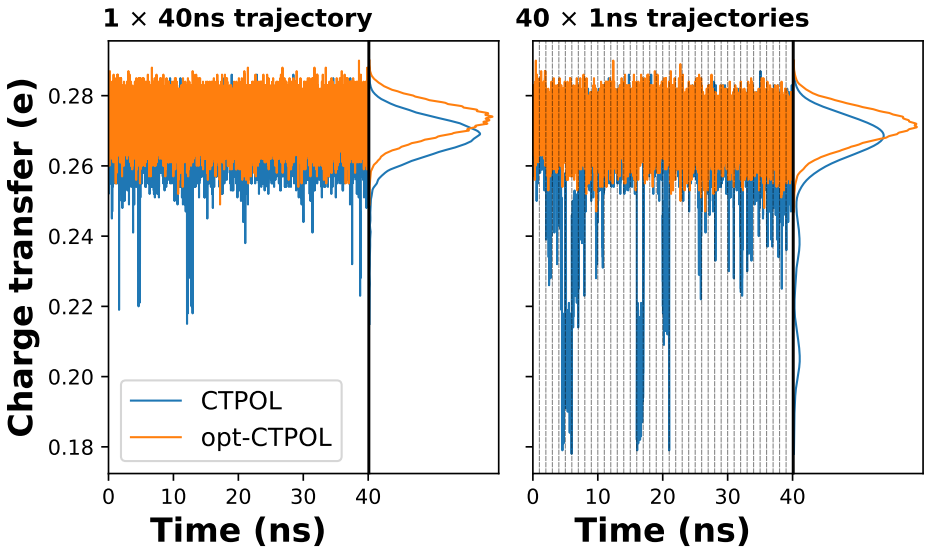


Figure 9: Charge transfer as a function of time for (left) a continuous 40 ns trajectory from one stable initial structure, and (right) 40 independent 1 ns simulations concatenated together. The dashed vertical lines mark the concatenation boundaries. The 40 × 1 ns simulations were started from different initial conditions randomly chosen from a continuous MD simulation, with randomized velocities.

down, the coordinating histidines containing these nitrogens move far away, as much as 9 Å away, but the sulfurs remain in close proximity at all times. At such distances, the charge transfer contribution of the nitrogens drop to zero, and the only contribution are from the sulfurs, and hence the lower total charge transfer. However, opt-CTPOL appears to have no such fluctuation in either the 40 ns or 40 × 1 ns trajectories.

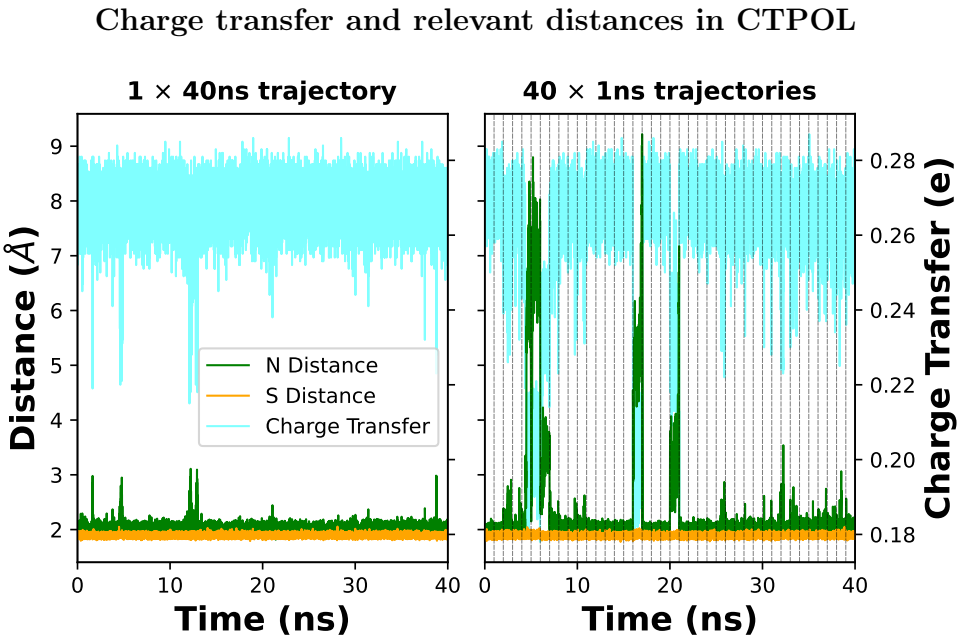
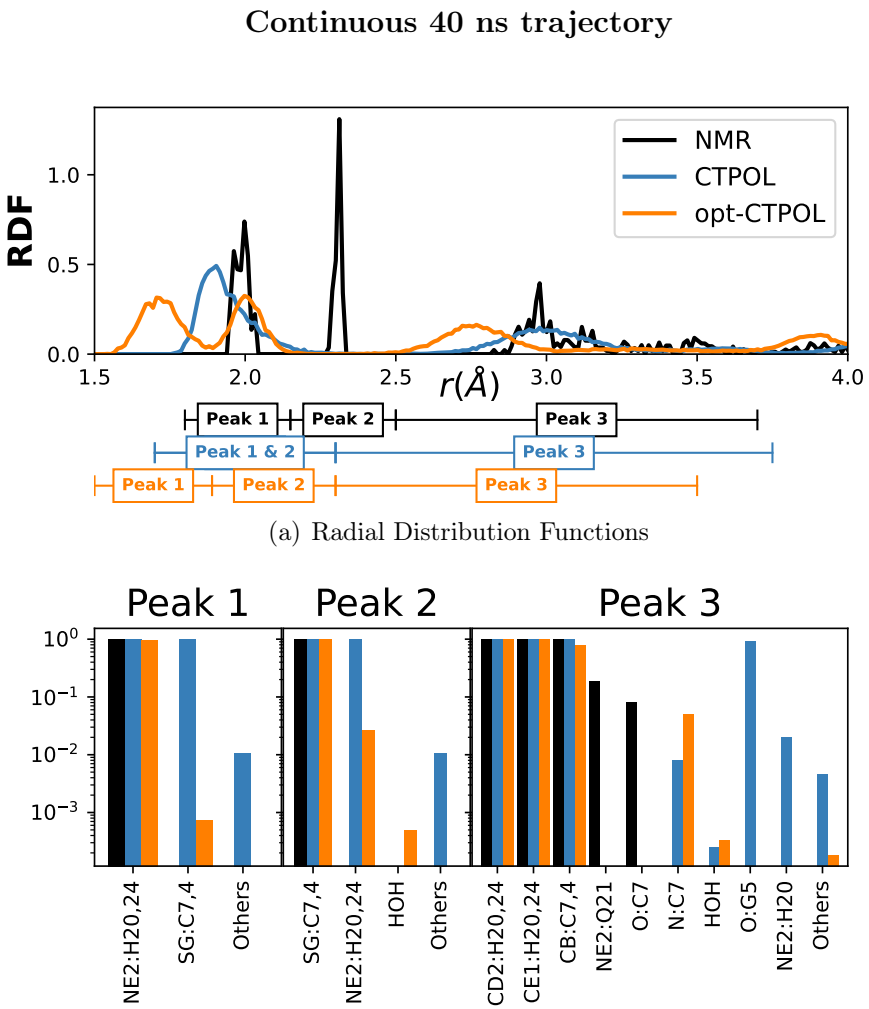


Figure 10: Coordinating nitrogen and sulfur distances (left y-axis) and charge transfer (right y-axis) vs time for a continuous trajectory (left) and 40 independent concatenated trajectories. In cyan, we have the charge transfer, in green, the average of the distances of Zn-N20 and Zn-N24, and in yellow the average of the distances of Zn-S4 and Zn-S7. Out of the 40 independent simulations, the average distance of Zn-N20/24 rises above 3 Å 8 times.

These unfolding events within 1ns occur about 20% of the time for CTPOL, thus making CTPOL without LJ-optimization unreliable.

Coordination structure and composition in opt-CTPOL shows improvement with a caveat.

To evaluate how parameters affect the coordination of Zn^{2+}, we plotted the radial distribution function of non-hydrogen protein atoms around the cation in Figure 11 (top).

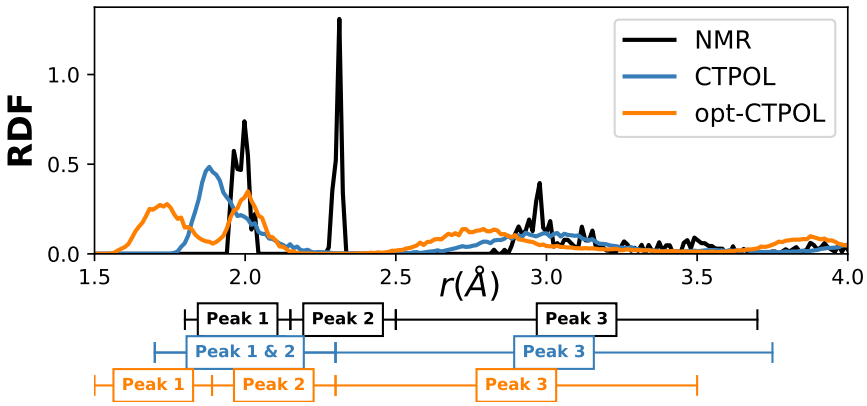


(a) Radial Distribution Functions

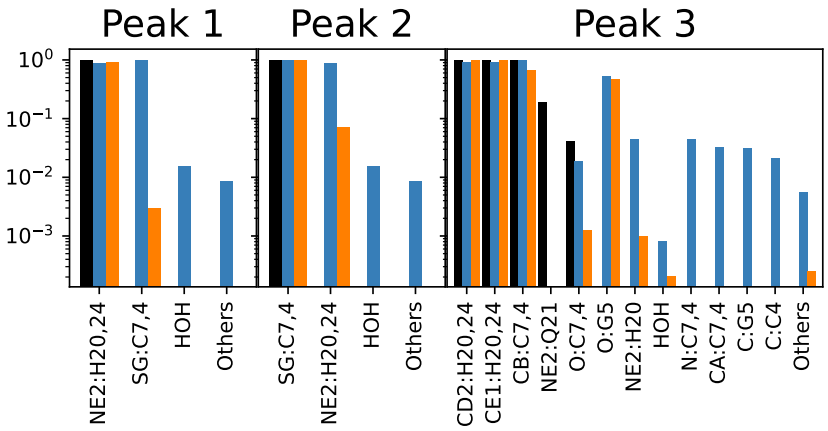
(b) Peak Breakdown

Figure 11: **Coordination analyses of continuous 40 ns trajectory.** a) RDF of all non-hydrogen protein atoms, with the distance ranges of selected peaks. b) Composition of each peak, where atoms of the same type and residue are lumped together. The Y-axis represents the average fraction of conformations in which each of the atoms appear within the peak range.

40×1**ns trajectory**



(a) Radial Distribution Functions



(b) Peak Breakdown

Figure 12: **Coordination analyses of** 40×1 **ns trajectory.** a) RDF of all non-hydrogen protein atoms, with the distance ranges of selected peaks. b) Composition of each peak, where atoms of the same type and residue are lumped together. The Y-axis represents the average fraction of conformations in which each of the atoms appear within the peak range.

We can see immediately that NMR and opt-CTPOL have a similar peak structure, but the distances are shorter in opt-CTPOL. In CTPOL, the first and second peaks, containing Nitrogens and Sulfurs respectively, overlap completely and are indistinguishable. This is not the case in the NMR models, where the Sulfur and Nitrogen peaks are quite distinct. In contrast to CTPOL, the opt-CTPOL peaks are distinct, with only a small percentage ($<$ 2%) of trajectories showing Nitrogens in the 2nd peak dominated by Sulfur. These features are also seen in similar analyses of the 40×1 ns trajectories (Figure 12). This is the first of a series of analyses in this paper that shows that CTPOL does not reproduce NMR binding domain as well as opt-CTPOL even for the stable continuous 40 ns trajectory.

After identifying the peaks, and selecting a range of distances (Figure 11 (top)), we determined which atoms comprise each peak and what fraction of the trajectory these atoms remain in that peak, as shown in Figure 11 (bottom). The 1st and 2nd peaks in CTPOL appear to be contaminated by other atom types which do not appear in NMR peaks at all. In the 40×1 ns trajectory, since CTPOL binding site has been shown to break apart in a few cases, it is no surprise that water also appears in Peak 1 of CTPOL (Figure 12 (bottom)). The opt-CTPOL model has no other atom-types in the first peak, and only relatively few others in the 2nd peak not present in NMR.

We should note that the NMR model we used does not contain any explicit water molecules. To determine if waters could be present in the binding site, we looked at 15 Zinc finger X-ray crystallography structures from the Protein Data Bank[113] (PDB) website (http://www.rcsb.org/pdb/) to find binding sites which are similar to this one (see S4 for a full list). We looked at binding sites which had a total of 2 histidines and 2 cysteines, similar to 1ZNF. We found 8 binding sites from the 15 crystal structures, and the smallest water distance to Zn^{2+} was 4.38 Å, well outside even the 3rd peak range in the NMR models. We further relaxed the matching criterion for the binding site to any binding site that contains a total of 4 histidines or cysteines (i.e., the number of coordinating histidines and cysteines sum to 4, but does not have to be 2 each). This resulted in a total of 60 binding sites. From

these, we found the smallest water distance to be 3.98 Å, still beyond the peak 3 range.

Thus, the inclusion of water in the 1st and 2nd peak, as is the case in CTPOL model, is uncharacteristic of Zn-finger binding sites of similar nature to 1ZNF. The opt-CTPOL model does a better job of keeping the water outside these peaks, with only a small fraction of waters in the 2nd peak.

Angle and distance distributions

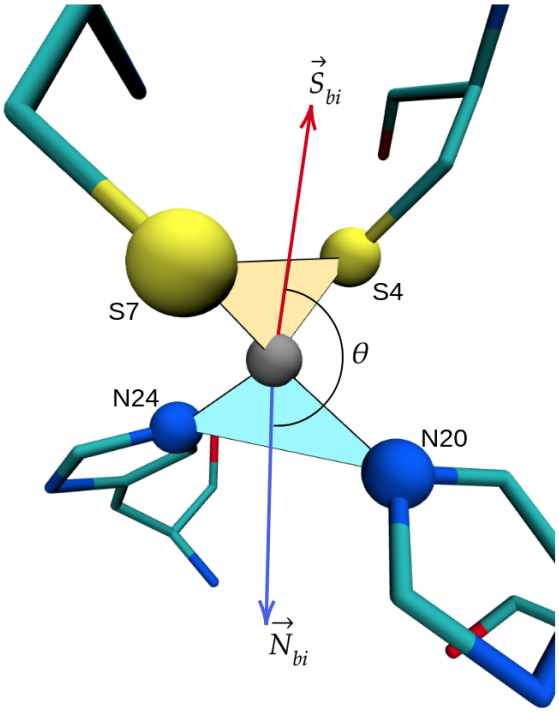


Figure 13: Binding site with Zn^{2+} at the center (grey atom), the sulfurs from Cys7 (S7) and Cys4 (S4), the NE nitrogens from His20 (N20) and His24 (24). Hydrogens have been removed for clarity. The yellow triangle on top has vertices on Zn, S4 and S7, while the blue triangle at the bottom has vertices on Zn, N20 and N24. The angle between the planes of these triangles are used for plotting the distributions in Figure 15. The red and blue arrows ($\vec{S}_{bi}$ and $\vec{N}_{bi}$) are vectors that bisect angles S7-Zn-S4 and N20-Zn-N24 respectively. The distributions of angle θ between these two bisectors are plotted on Figure 15 (b). The distributions of some of the distances between the 5 atoms shown in this figure are shown on Figure 16, while the distributions for some of the angles are shown on Figure 14.

To further evaluate the stability and accuracy of the binding domain in the CTPOL and opt-CTPOL frameworks, we analyzed a number of geometric quantities which are defined in Figure 13 and its caption. Here we only consider the 40 ns continuous trajectory for which the binding domain is stable for CTPOL, since these geometric quantities would not make sense for the 40 × 1ns trajectory where the binding domain destabilizes.

Figure 14 shows the distribution of most of the angles that the coordinating atoms make with Zn^{2+}. Additionally, Figure 15 (a) shows the distributions of angles between the planes

shown in Figure 13, and Figure 15 (b) shows the distributions of the angles between the bisectors, also defined in Figure 13. It is quite clear that opt-CTPOL reproduces the NMR distributions of angles as well or better than CTPOL. The distribution of the S4-Zn-S7 angle appears to agree particularly well with NMR, as does the angle between the bisectors. While the CTPOL 40 ns trajectory showed a slightly better overall RMSD from Figure 6, it is clearly not reproducing these angles as well as opt-CTPOL. This implies that opt-CTPOL is maintaining the shape of the binding domain better, which is in accordance with the RDF distribution and peak analysis of Figure 11.

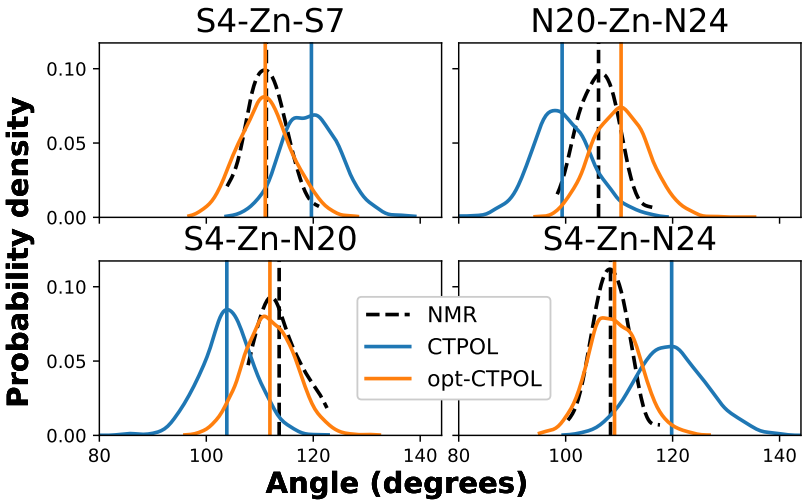
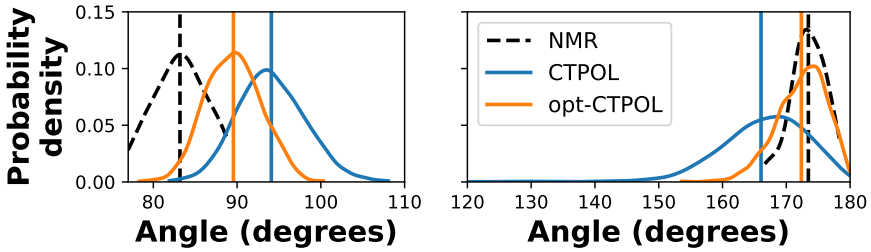


Figure 14: Probability distribution of angles over the continuous 40 ns trajectories of CTPOL (blue) and opt-CTPOL (orange) and over 37 NMR models (black dashed). The corresponding atoms are depicted in Figure 13. The distributions were calculated using Kernel Density Estimation.[110,111] The vertical lines represent the averages of each distribution.

Furthermore, we see from Figure 16 that the distances of opt-CTPOL binding domain are consistently shorter than those of the experimental NMR structures. This is in line with the RDF analysis of Figure 11, where we see similar peak structure of opt-CTPOL, but at shorter distances. On the other hand, CTPOL distances do not appear to have a consistent relation to the NMR distances. For instance, the distances of S*-Zn and N*-Zn (top left) show that opt-CTPOL distances trend the same way as NMR, i.e., the N*-Zn distances are significantly shorter than S*-Zn distances. For CTPOL, it turns out to be almost the



(a) Dihedral distribution (b) Bisector angle distribution

Figure 15: (a) Probability distributions of angle between the S7-Zn-S4 and N24-Zn-N20 planes as depicted in Figure 13. (b) Angle between S4-Zn-S7 and N20-Zn-N24 bisectors, which are depicted in Figure 13 as $\vec{S}_{bi}$ and $\vec{N}_{bi}$, respectively.

opposite, with plenty of overlap between the two distributions, and thus their 1st and 2nd peaks in Figure 11 also overlap.

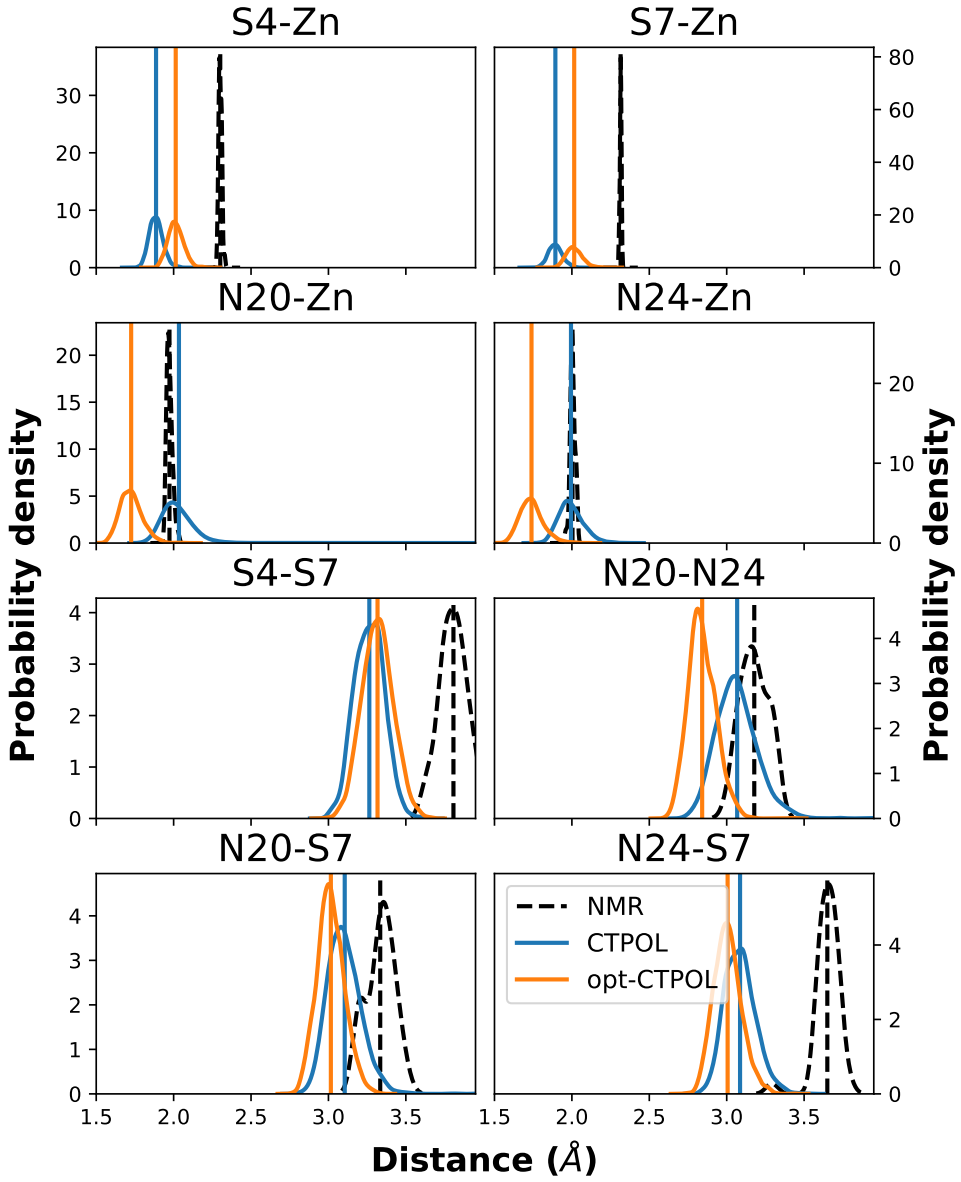


Figure 16: Probability distribution of distances (using Kernel Density Estimation[110,111]) over entire trajectory (for simulations) and over 37 models (for NMR data).

4 Conclusion and outlook

The development of accurate force field parameters for cation-peptide systems is a major obstacle in metalloprotein simulations. One approach to facilitate the development of new force field parameters is to construct tools to derive parameters from QM calculations. Our previous work[21] has shown that QM-driven parameterization of Drude and CTPOL models may improve the accuracy of the description of ion-protein interactions in MD simulations. However, the Drude model may be limited when charge transfer effects are significant. Furthermore, the additional particles attached to polarizable atoms by the Drude model are light. To capture the vibrations of these light particles, the time step of the Drude model must be small, which makes the Drude model computationally more expensive than the CTPOL model.

In this regard, FFAFFURR is developed as a python tool to facilitate the parameterization of classical and polarizable CTPOL models. In this paper, we chose to parameterize OPLS-AA as an example. However, the tool should also work with other similar force fields such as CHARMM and AMBER once the code for parsing parameters is generalized.

QM calculations from FHI-aims can be automatically parsed to FFAFFURR, and the output parameter files of FFAFURR can be directly processed by the molecular dynamic package OpenMM. All energy terms in OPLS-AA and CTPOL models can be tuned by FFAFFURR. The performance of optimized parameters in each energy term was evaluated by the comparison of FF energies and QM potential energies. Users can choose which energy term to adjust in practice. We showed that the CTPOL model outperforms OPLS-AA in terms of QM energy reproduction for divalent-dipeptide systems.

One potential usage of FFAFFURR is the rapid construction of FFs for troublesome metal centers in metalloproteins. We tested this function by performing MD simulations on the 1ZNF Zinc finger protein[102] and comparing simulation results with NMR models. With the parameters optimized from FFAFFURR, we found that both CTPOL and opt-CTPOL better reproduce the overall structure of the protein. However, to better stabilize

and reproduce structural features of the binding domain, LJ optimization (opt-CTPOL) was necessary, since CTPOL alone had some shortcomings in correctly reproducing the binding domain, or keeping it stable under various initial conditions. The LJ optimization resulted in coordination composition and geometry that better agrees with the NMR models than CTPOL alone. On the other hand, the optimization of LJ does lead to a somewhat shrunken binding domain. Whether this is a major concern remains to be seen with further studies, such as calculations of relative binding affinities with other metals, or other macroscopic analyses of similar systems which could be verified experimentally.

In summary, FFAFFURR has a wide range of functions and can provide almost all the functions required for the cation-peptide parameterization process. FFAFFURR helps users to get rid of labor-intensive steps in FF optimization.

Despite the success of FFAFFURR in this study, we see several directions to discuss in future research. Note that only the parameters of the Zinc finger protein interaction center were optimized with FFAFFURR in the MD simulation, while the standard OPLS-AA parameters were used for the rest of the protein. While our study indicates the compatibility of the optimized parameters with the standard FF parameters, this may need to be investigated in more detail in a future study. One characteristic of FFAFFURR is that it can be employed to derive parameters for a specific system. This helps to grasp the specific environment of the system. However, QM calculations are required when a new system is under investigation. We created a data set of cation-dipeptides containing several divalent cations, which can be automatically parsed to FFAFFURR.[67] If the user's system goes beyond the scope of the dataset, an in-house genetic algorithm package Fafoom[87] can be used to generate conformers and do the QM generation fast and automatically.

Acknowledgement

The authors would like to thank:

The China Scholarship Council for providing X.H. with a doctoral fellowship;

The Federal Ministry of Education and Research of Germany for providing funding for the project STREAM ("Semantische Repräsentation, Vernetzung und 333 Kuratierung von qualitätsgesicherten Materialdaten", ID: 16QK11C); and

MITACS for the MITACS Globalink Research Award which funded the visit of K.S.A. to the lab of X.H. and C.B.

References

(1) Sarkar, B. Metal protein interactions. *Prog. Food Nutr. Sci.* **1987**, *11*, 363–400.

(2) Peters, M. B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M. N.; Merz, K. M. Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *J. Chem. Theory Comput.* **2010**, *6*, 2935–2947.

(3) Christianson, D. W. Structural biology of zinc. *Adv. Protein Chem.* **1991**, *42*, 281–355.

(4) Patel, K.; Kumar, A.; Durani, S. Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures. *Biochim. Biophys. Acta - Proteins Proteom.* **2007**, *1774*, 1247–1253.

(5) Babu, C. S.; Lee, Y. M.; Dudev, T.; Lim, C. Modeling Zn^{2+} release from metallothionein. *J. Phys. Chem. A* **2014**, *118*, 9244–9252.

(6) Bell, S. G.; Vallee, B. L. The metallothionein/thionein system: An oxidoreductive metabolic zinc link. *ChemBioChem* **2009**, *10*, 55–62.

(7) Capdevila, M.; Bofill, R.; Palacios, O.; Atrian, S. State-of-the-art of metallothioneins at the beginning of the 21st century. *Coord. Chem. Rev.* **2012**, *256*, 46–62.

(8) Cherian, G.; Jayasurya, A.; Bay, B.-H. Metallothionein in human tumors and potential carcinogenesis. *Mutat. Res.* **2004**, *533*, 201–9.

(9) Durand, J.; Meloni, G.; Talmard, C.; Vašák, M.; Faller, P. Zinc release of Zn_7-metallothionein-3 induces fibrillar type amyloid-β aggregates. *Metallomics* **2010**, *2*, 741–744.

(10) Laity, J. H.; Lee, B. M.; Wright, P. E. Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* **2001**, *11*, 39–46.

(11) Lipscomb, W. N.; Sträter, N. Recent advances in zinc enzymology. *Chem. Rev.* **1996**, *96*, 2375–2434.

(12) Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.* **2012**, *137*, 230901.

(13) Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D. An empirical polarizable force field based on the classical drude oscillator model: Development history and recent applications. *Chem. Rev.* **2016**, *116*, 4983–5013.

(14) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813 – 1818.

(15) A. Kaminski, G.; A. Friesner, R.; Tirado-Rives, J.; L. Jorgensen, W. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(16) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198–210.

(17) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73.

(18) Reif, M. M.; Winger, M.; Oostenbrink, C. Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins. *J. Chem. Theory Comput.* **2013**, *9*, 1247–1264.

(19) Li, H.; Ngo, V.; Da Silva, M. C.; Salahub, D. R.; Callahan, K.; Roux, B.; Noskov, S. Y. Representation of ion-protein interactions using the Drude polarizable force-field. *J. Phys. Chem. B* **2015**, *119*, 9401–9416.

(20) Li, P.; Merz, K. M. Metal ion modeling using classical mechanics. *Chem. Rev.* **2017**, *117*, 1564–1686.

(21) Amin, K. S.; Hu, X.; Salahub, D. R.; Baldauf, C.; Lim, C.; Noskov, S. Benchmarking polarizable and non-polarizable force fields for Ca^{2+}–peptides against a comprehensive QM dataset. *J. Chem. Phys.* **2020**, *153*, 144102.

(22) Maksimov, D.; Baldauf, C.; Rossi, M. The conformational space of a flexible amino acid at metallic surfaces. *Int. J. Quantum Chem.* **2021**, *121*, e26369.

(23) Schneider, M.; Baldauf, C. Relative energetics of acetyl-histidine protomers with and without Zn^{2+} and a benchmark of energy methods. *arXiv preprint arXiv:1810.10596* **2018**,

(24) Wu, J. C.; Piquemal, J.-P.; Chaudret, R.; Reinhardt, P.; Ren, P. Polarizable molecular dynamics simulation of Zn (II) in water using the AMOEBA force field. *J. Chem. Theory Comput.* **2010**, *6*, 2059–2070.

(25) Akin-Ojo, O.; Song, Y.; Wang, F. Developing ab initio quality force fields from

condensed phase quantum-mechanics/molecular-mechanics calculations through the adaptive force matching method. *J. Chem. Phys.* **2008**, *129*, 64108.

(26) Duboué-Dijon, E.; Javanainen, M.; Delcroix, P.; Jungwirth, P.; Martinez-Seara, H. A practical guide to biologically relevant molecular simulations with charge scaling for electronic polarization. *J. Chem. Phys.* **2020**, *153*, 50901.

(27) Martinek, T.; Duboué-Dijon, E.; Timr, Š.; Mason, P. E.; Baxová, K.; Fischer, H. E.; Schmidt, B.; Pluhařová, E.; Jungwirth, P. Calcium ions in aqueous solutions: Accurate force field description aided by ab initio molecular dynamics and neutron scattering. *J. Chem. Phys.* **2018**, *148*, 222813.

(28) Le Breton, G.; Joly, L. Molecular modeling of aqueous electrolytes at interfaces: Effects of long-range dispersion forces and of ionic charge rescaling. *J. Chem. Phys.* **2020**, *152*, 241102.

(29) Li, P.; Song, L. F.; Merz Jr, K. M. Systematic parameterization of monovalent ions employing the nonbonded model. *J. Chem. Theory Comput.* **2015**, *11*, 1645–1657.

(30) Li, P.; Song, L. F.; Merz Jr, K. M. Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water. *J. Phys. Chem. B* **2015**, *119*, 883–895.

(31) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr., A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(32) Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336.

(33) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE-Antechamber python parser interface. *BMC Res. Notes* **2012**, *5*, 1–8.

(34) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(35) Vassetti, D.; Pagliai, M.; Procacci, P. Assessment of GAFF2 and OPLS-AA general force fields in combination with the water models TIP3P, SPCE, and OPC3 for the solvation free energy of druglike organic molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1983–1995.

(36) Kumar, A.; Yoluk, O.; MacKerell Jr, A. D. FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules. *J. Comput. Chem.* **2020**, *41*, 958–970.

(37) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building force fields: An automatic, systematic, and reproducible approach. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

(38) Jorgensen, W. L.; Jensen, K. P.; Alexandrova, A. N. Polarization effects for hydrogen-bonded complexes of substituted phenols with water and chloride ion. *J. Chem. Theory Comput.* **2007**, *3*, 1987–1992.

(39) Tkatchenko, A.; DiStasio Jr, R. A.; Car, R.; Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

(40) Tkatchenko, A.; Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **2009**, *102*, 6–9.

(41) Gobre, V. V.; Tkatchenko, A. Scaling laws for van der Waals interactions in nanostructured materials. *Nat. Commun.* **2013**, *4*, 2341.

(42) Horton, J. T.; Allen, A. E.; Dodda, L. S.; Cole, D. J. QUBEKit: Automating the derivation of force field parameters from quantum mechanics. *J. Chem. Inf. Model.* **2019**, *59*, 1366–1381.

(43) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular force field parameterization via atoms-in-molecule electron density partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312–2323.

(44) Grimme, S. A general quantum mechanically derived force field (QMDFF) for molecules and condensed phase simulations. *J. Chem. Theory Comput.* **2014**, *10*, 4497–4514.

(45) Borodin, O. Polarizable force field development and molecular dynamics simulations of ionic liquids. *J. Phys. Chem. B* **2009**, *113*, 11463–11478.

(46) Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. Polarization effects in molecular mechanical force fields. *J. Phys. Condens. Matter* **2009**, *21*, 333102.

(47) Allen, T. W.; Andersen, O. S.; Roux, B. Energetics of ion conduction through the gramicidin channel. *Proc. Natl. Acad. Sci.* **2004**, *101*, 117 – 122.

(48) Boulanger, E.; Thiel, W. Toward QM/MM simulation of enzymatic reactions with the Drude oscillator polarizable force field. *J. Chem. Theory Comput.* **2014**, *10*, 1795–1809.

(49) Panel, N.; Villa, F.; Fuentes, E. J.; Simonson, T. Accurate PDZ/peptide binding specificity with additive and polarizable free energy simulations. *Biophys. J.* **2018**, *114*, 1091–1102.

(50) Li, Y. L.; Mei, Y.; Zhang, D. W.; Xie, D. Q.; Zhang, J. Z. H. Structure and dynamics of a dizinc metalloprotein: effect of charge transfer and polarization. *J. Phys. Chem. B* **2011**, *115*, 10154–10162.

(51) Bedrov, D.; Piquemal, J.-P.; Borodin, O.; MacKerell, A. D.; Roux, B.; Schröder, C. Molecular dynamics simulations of ionic liquids and electrolytes using polarizable force fields. *Chem. Rev.* **2019**, *119*, 7940–7995.

(52) Olano, L. R.; Rick, S. W. Fluctuating charge normal modes: An algorithm for implementing molecular dynamics simulations with polarizable potentials. *J. Comput. Chem.* **2005**, *26*, 699–707.

(53) Soniat, M.; Rick, S. W. The effects of charge transfer on the aqueous solvation of ions. *J. Chem. Phys.* **2012**, *137*, 044511.

(54) Piquemal, J.-P.; Chevreau, H.; Gresh, N. Toward a separate reproduction of the contributions to the Hartree-Fock and DFT intermolecular interaction energies by polarizable molecular mechanics with the SIBFA potential. *J. Chem. Theory Comput.* **2007**, *3*, 824–837.

(55) Friesner, R. A. Modeling polarization in proteins and protein–ligand complexes: Methods and preliminary results. *Adv Protein Chem.* **2005**, *72*, 79–104.

(56) Cieplak, P.; Caldwell, J.; Kollman, P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J. Comput. Chem.* **2001**, *22*, 1048–1057.

(57) Ponder, J. W.; Case, D. A. *Advances in protein chemistry*; Elsevier, 2003; Vol. 66; pp 27–85.

(58) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(59) Ngo, V.; da Silva, M. C.; Kubillus, M.; Li, H.; Roux, B.; Elstner, M.; Cui, Q.; Salahub, D. R.; Noskov, S. Y. Quantum effects in cation interactions with first and second coordination shell ligands in metalloproteins. *J. Chem. Theory Comput.* **2015**, *11*, 4992–5001.

(60) Villa, F.; MacKerell Jr, A. D.; Roux, B.; Simonson, T. Classical Drude polarizable force field model for methyl phosphate and its interactions with Mg^{2+}. *J. Phys. Chem. A* **2018**, *122*, 6147–6155.

(61) Dudev, T.; Lim, C. Competition among metal ions for protein binding sites: Determinants of metal ion selectivity in proteins. *Chem. Rev.* **2014**, *114*, 538–556.

(62) Ngo, V.; da Silva, M. C.; Kubillus, M.; Li, H.; Roux, B.; Elstner, M.; Cui, Q.; Salahub, D. R.; Noskov, S. Y. Quantum effects in cation interactions with first and second coordination shell ligands in metalloproteins. *J. Chem. Theory Comput.* **2015**, *11*, 4992–5001.

(63) Dudev, T.; Lin, Y.-l.; Dudev, M.; Lim, C. First-second shell interactions in metal binding sites in proteins: A PDB survey and DFT/CDM calculations. *J. Am. Chem. Soc.* **2003**, *125*, 3168–3180.

(64) Sakharov, D. V.; Lim, C. Zn protein simulations including charge transfer and local polarization effects. *J. Am. Chem. Soc.* **2005**, *127*, 4921–4929.

(65) Sakharov, D. V.; Lim, C. Force fields including charge transfer and local polarization effects: Application to proteins containing multi/heavy metal ions. *J. Comput. Chem.* **2009**, *30*, 191–202.

(66) Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* **2016**, *3*, 1–13.

(67) Hu, X.; Lenz-Himmer, M.-O.; Baldauf, C. Better force fields start with better data: A data set of cation dipeptide interactions. *Sci. Data* **2022**, *9*, 1–14.

(68) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

(69) Hu, X.; Baldauf, C. First release of CTPOL_MD code. *zenodo* **2023**,

(70) Hu, X.; Baldauf, C. FFAFFURR Data set. *NOMAD* **2023**,

(71) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.

(72) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. Efficient O(N) integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **2009**, *228*, 8367–8379.

(73) Ren, X.; Rinke, P.; Blum, V.; Wieferink, J.; Tkatchenko, A.; Sanfilippo, A.; Reuter, K.; Scheffler, M. Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New J. Phys.* **2012**, *14*, 053020.

(74) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(75) Tkatchenko, A.; Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

(76) Ropo, M.; Blum, V.; Baldauf, C. Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead. *Sci. Rep.* **2016**, *6*, 1–11.

(77) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbó-Dorca, R. Critical analysis and extension of the Hirshfeld atoms in molecules. *J. Chem. Phys.* **2007**, *126*.

(78) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129–138.

(79) Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129.

(80) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(81) Rossi, M.; Chutia, S.; Scheffler, M.; Blum, V. Validation challenge of density-functional theory for peptides-example of Ac-Phe-Ala_5-$LysH^+$. *J. Phys. Chem. A* **2014**, *118*, 7349–7359.

(82) Wales, D. J.; K., D. P. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

(83) Wales, D. J.; Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **1999**, *285*, 1368–1372.

(84) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(85) Ponder, J. W.; Richards, F. M. An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.* **1987**, *8*, 1016–1024.

(86) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(87) Supady, A.; Blum, V.; Baldauf, C. First-principles molecular structure search with a genetic algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2338–2348.

(88) Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288.

(89) Hoerl, A. E.; Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67.

(90) Shi, Y.; Eberhart, R. C. Empirical study of particle swarm optimization. Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406). 1999; pp 1945–1950.

(91) Koh, B.-I.; George, A. D.; Haftka, R. T.; Fregly, B. J. Parallel asynchronous particle swarm optimization. *Int. J. Numer. Methods Eng.* **2006**, *67*, 578–595.

(92) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(93) Lee, A. 2014; https://pythonhosted.org/pyswarm/.

(94) Dittner, M.; Müller, J.; Aktulga, H. M.; Hartke, B. Efficient global optimization of reactive force-field parameters. *J. Comput. Chem.* **2015**, *36*, 1550–1561.

(95) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S. Building a more predictive

protein force field: a systematic and reproducible route to AMBER-FB15. *J. Phys. Chem. B* **2017**, *121*, 4023–4039.

(96) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. Systematic ab initio gradient calculation of molecular geometries, force constants, and dipole moment derivatives. *J. Am. Chem. Soc.* **1979**, *101*, 2550–2560.

(97) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid parameterization of small molecules using the force field toolkit. *J. Comput. Chem.* **2013**, *34*, 2757–2770.

(98) Klimeš, J.; Michaelides, A. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.* **2012**, *137*, 120901.

(99) Reilly, A. M.; Tkatchenko, A. van der Waals dispersion interactions in molecular materials: beyond pairwise additivity. *Chem. Sci.* **2015**, *6*, 3289–3301.

(100) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular force field parameterization via atoms-in-molecule electron density partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312–2323.

(101) Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(102) Lee, M. S.; Gippert, G. P.; Soman, K. V.; Case, D. A.; Wright, P. E. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **1989**, *245*, 635–637.

(103) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(104) Zhang, J.; Yang, W.; Piquemal, J.-P.; Ren, P. Modeling structural coordination and

ligand binding in zinc proteins with a polarizable potential. *J. Chem. Theory Comput.* **2012**, *8*, 1314–1324.

(105) Baldauf, C.; Pagel, K.; Warnke, S.; Von Helden, G.; Koksch, B.; Blum, V.; Scheffler, M. How cations change peptide structure. *Chem. - Eur. J.* **2013**, *19*, 11224–11234.

(106) Miller, J.; McLachlan, A. D.; Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J.* **1985**, *4*, 1609–1614.

(107) Wolfe, S. A.; Nekludova, L.; Pabo, C. O. DNA recognition by (Cys_2His_2) zinc finger proteins. *Annu. Rev. Bioph. Biom.* **2000**, *29*, 183.

(108) Gamsjaeger, R.; Liew, C. K.; Loughlin, F. E.; Crossley, M.; Mackay, J. P. Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends Biochem. Sci.* **2007**, *32*, 63–70.

(109) Li, P.; Merz Jr, K. M. MCPB.py: a python based metal center parameter builder. *J. Chem. Inf. Model.* **2016**, *56*, 599–604.

(110) Parzen, E. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **1962**, *33*, 1065 – 1076.

(111) Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* **1956**, *27*, 832 – 837.

(112) Donini, O. A.; Kollman, P. A. Calculation and prediction of binding free energies for the matrix metalloproteinases. *J. Med. Chem.* **2000**, *43*, 4180–4188.

(113) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

Supporting Information Available

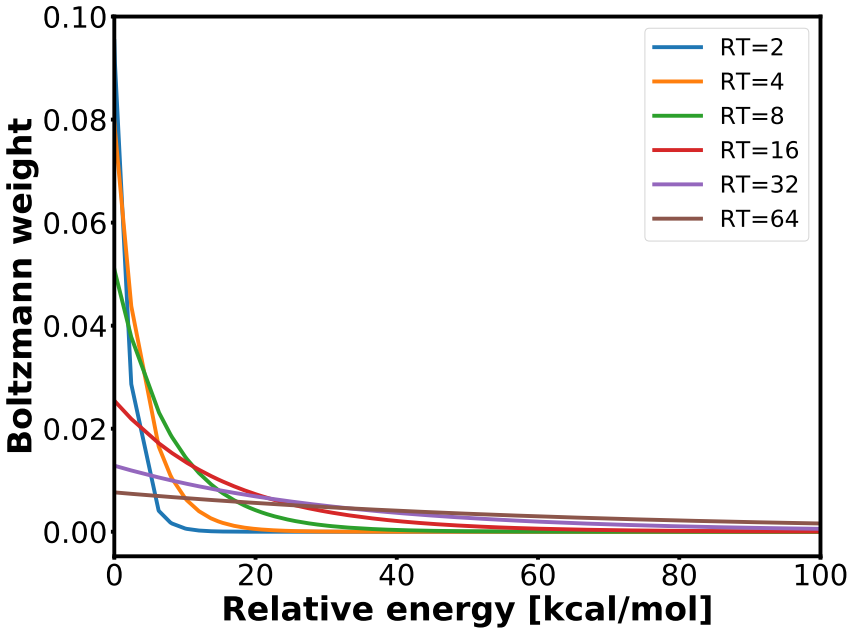


Figure S1: Boltzmann-type weights vs. relative QM energies at various RTs of AcCys^{-}NMe+Zn^{2+} system.

Table S1: Atom types in HisD+Zn^{2+} and Cys^{-}+Zn^{2+}.

HisD+Zn^{2+}		Cys^{-}+Zn^{2+}	
Atom	Atom type	Atom	Atom type
C	2177	C	1177
CA	2166	CA	1166
CB	2446	CB	1148
CD2	2448	H	1183
CE1	2447	HA	1086
CG	2449	HB2	1085
H	2183	HB3	1085
HA	2086	N	1180
HB2	2085	O	1178
HB3	2085	SG	1142
HD1	2445		
HD2	2091		
HE1	2092		
N	2180		
ND1	2444		
NE2	2452		
O	2178		
Zn	834		

Table S2: The optimized LJ parameters. Epsilon = 0 means the LJ interaction is neglected. LASSO tends to focus on only important factors while neglecting insignificant ones.

Type1	Type2	Sigma (nm)	Epsilon (kJ/mol)
2178	834	0.31933	0.00024413
2448	834	0.331094	0
2183	834	0.32642	0.001277
2446	834	0.32934	0.03138
2177	834	0.330867	0
2092	834	0.288564	0
2091	834	0.288564	0
2180	834	0.319954	0
2447	834	0.329767	0
2444	834	0.31992	0.25885
2445	834	0.29663	6.7250
2085	834	0.294726	0
2086	834	0.294726	0
2184	834	0.325209	0
2452	834	0.32598	0.00153
2166	834	0.331252	0.01205
2449	834	0.33125	0.01205

Table S3: The CTPOL parameters. The a and b are parameters in eq. 11, r is the cutoff distance. The correction factor k in eq. 11 is set as 3.418.

Type	Polarizability (nm^3)	a	b	r (nm)
1142	0.002668	-1.037	0.323	0.312
1178	0.000729	-0.246	0.072	0.294
1180	0.00093	-0.478	0.129	0.270
2178	0.000721	-2.667	0.722	0.271
2180	0.000901	-0.635	0.172	0.270
2452	0.000952	-0.593	0.193	0.325
2444	0.000879	-2.424	0.843	0.348
444	0.000879			
452	0.000952			
834	0.004383			
166/2166/1166	0.001454			
447/2447	0.001341			
448/2448	0.001416			
80	0.001316			
1177	0.001473			
446/2446	0.001397			
177/2177	0.001441			
178	0.000721			
184	0.001292			
449/2449	0.001446			
180	0.000901			
1148	0.001475			
96	0.000724			
250	0.001394			
246	0.000906			
235	0.001339			
81/82	0.001431			
243	0.000864			
94	0.001457			
108	0.001497			
214	0.000858			
213	0.001685			
230	0.000810			
179	0.000904			
165	0.001425			
90	0.001471			
251	0.001417			
216	0.001504			
109	0.000711			
99	0.001439			
245	0.001410			

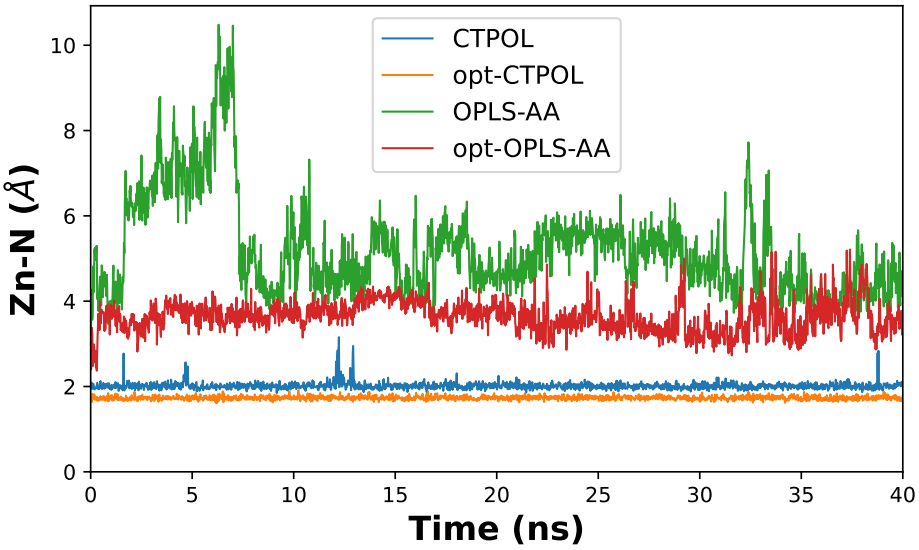


Figure S2: Average of the two Zn-N distances, where the N are the NE2 atoms of the two histidines in the binding site, as a function of time.

Table S4: PDB ids of Zn fingers. N4HC denotes number of Zn binding sites with 4 His and Cys residues, whereas N2H2C denotes the number of Zn binding sites with exactly 2 His and 2 Cys. The last column denotes the distance of the closest water molecule to the Zn ion.

PDBid	Zn_sites	N4HC	N2H2C	Min H2O dist
1MEY	8	7	7	4.38
4QF3	4	4	0	3.98
6UEI	4	4	0	4.24
6UEJ	4	4	0	4.30
2PUY	4	4	0	4.35
6FI1	4	4	0	9.00
6FHQ	4	4	0	4.04
5YC3	2	2	0	6.49
3T7L	2	2	0	4.41
3U9G	4	4	0	4.26
4Q6F	8	8	0	4.32
3IUF	1	1	1	5.42
4BBQ	8	8	0	4.44
5YC4	2	2	0	6.60
5Y20	2	2	0	5.66

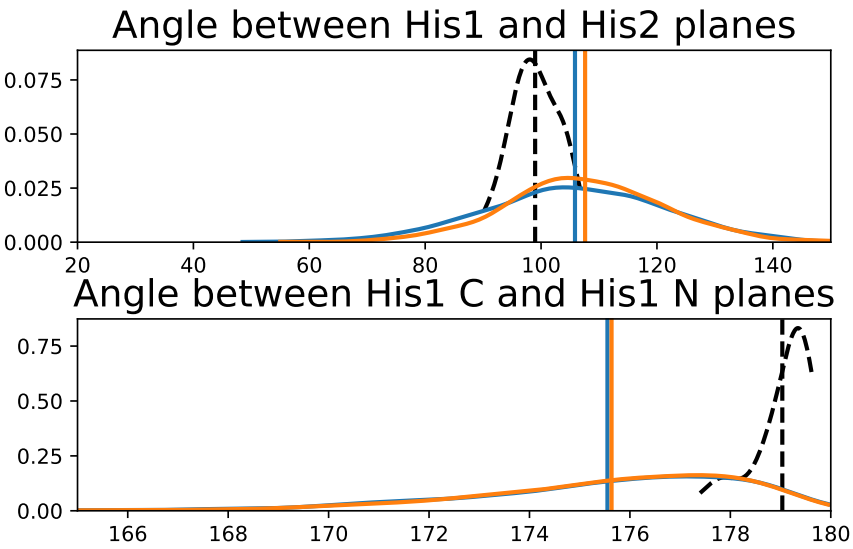


Figure S3: Probability distributions of select dihedral angles. (top) The dihedral angle between the two coordinating histidine planes. The planes were determined using the CG, CD, and CE atoms of histidine. (bottom) The dihedral angle between plane defined by His1 CG, CD, and CE1 atoms, and plane defined by His1 CG, ND, and NE atoms. This is to check for internal distortion of the plane. The values are close to 180 (instead of 0) because one set of atoms goes clockwise, and the other counter clockwise, when defining the planes.

Chapter 5

Conclusions and Outlook

In this thesis, I study force field development for cation-dipeptide systems. The simulation of metalloproteins is a long-standing challenging topic. Bonded force field models of the cation binding site succeed in maintaining the correct coordination number of the protein interaction center by employing covalent bonds between cations and ligands atoms. However, cations are artificially fixed at the interaction center, so the bonded models can not simulate the formation and disassembly of metalloproteins or the transmembrane passage of ions. With the three manuscripts that make up this thesis, I contribute to the development of the next-generation non-bonded force field models for cation-dipeptides simulations:

- The first part of this thesis provides a large and comprehensive quantum chemistry data set of cation-dipeptides as a solid database for force field development or benchmarking. The representation by an ontology makes the data set fulfill the FAIR principles: findable, accessible, interoperable, and re-usable.

- The second part of the thesis benchmarks three popular non-polarizable FFs: CHARMM36, AMBER10, OPLS-AA, as well as a polarizable Drude model using the QM data set. The result shows that the polarizable Drude model performs better in reproducing the DFT interaction energies than non-polarizable models. However, several Thole factors and LJ parameters need to be parameterized to ameliorate the polarization catastrophe. The optimized parameters were validated by MD simulation of a metalloprotein. Since the Drude model can be limited in case of significant charge transfer effects, the CTPOL model that explicitly takes charge transfer and polarization effects into account was also tested.

- The third part of the thesis implements the CTPOL model and provides an open source python tool FFAFFURR that enables the parameterization of OPLS-AA and CTPOL model. The parameters of a energy terms in the OPLS-AA and CTPOL models can be adjusted by FFAFFURR. The performance of the parameterized parameters was evaluated by their ability to reproduce DFT energies and MD simulations of zinc finger protein in solution.

The details of the studies have been summarized in Chapter 3, but there are some points I

would like to highlight:

- A comprehensive and accurate database is essential to FFs development.
- With the development of computational physics and chemistry, the amount of QM data is rapidly increasing. How to share this data becomes more and more important. Ontologies can link the data in a formalized machine-understandable way. Although it takes time and effort to create a comprehensive ontology, ontologies have great potential to organize huge amounts of data and serve as the knowledge structure behind AI technologies.
- The polarizable FF models can better simulate metalloproteins than non-polarizable FFs. However, some parameters of polarizable FF models may need to be re-parameterized and sometimes specified for the system of interest.
- The parameters of FFs are interconnected to each other. Adjusting the parameters of one energy term may affect the parameters of other energy terms. Therefore, some energy terms may need to be adjusted jointly. FFAFFURR enables the parameterization of all energy terms. Users can choose the parameters to tune according to the actual application.
- We perform FF optimization based on molecular/atomic properties (e.g., partial charge), and fitting to only energies, but no forces.
- QM-driven parameterization combined with MD simulations for testing is a good way to develop the FFs for metalloproteins.
- In addition to polarization effects, the CTPOL model explicitly includes charge transfer effects. Moreover, the CTPOL model does not, in contrast to Drude model, add light particles to the model, which allows simulations with the CTPOL model to take larger time steps. These make the CTPOL model a promising method to accurately describe metalloproteins with relatively low computational costs.
- Re-parameterization is tedious and time consuming. User friendly parameterization tools can greatly facilitate FF development.

In summary, this thesis contributes to the development of cation-dipeptide FFs from four aspects: (1) providing a comprehensively quantum chemistry data set of cation-dipeptides, (2) investigating methods for QM-driven FF development and validation, (3) implementing CTPOL model, and (4) providing free and open-source FFs parameterization tool.

Although many advances have been made in this thesis, there is still much room for further exploration. Some ideas are listed below:

- The ontologies and knowledge graph developed in this thesis makes the data set linked. New knowledge, relationships, or trends in the data set may be found through the investigation of the data set by SPARQL queries. This will deepen the understanding of the cation-dipeptide systems. The ontologies can also be linked to external ontologies to connect knowledge across projects and fields.

- The quantum chemistry data set is organized as knowledge graph. The data as well as metadata can be accessed through the SPARQL query language. It is reasonable to implement an interface between the knowledge graph and FFAFFURR. This will lead to easy access to the quantum chemistry data and speed up the parameterization process. It also makes it easy to extend the quantum chemistry data by adding new data to the knowledge graph.
- There are several groups focusing on the development of better Drude models. It would be interesting to compare the performance of the optimized Drude model and the CTPOL model in metalloprotein simulations.
- Introducing more target properties, e.g. binding energy, may improve the predictive power of the new parameter set.

Appendix

A Software for ontology and knowledge graph development

Protégé

Protégé is a free, and open-source graphical program for editing and exploring ontologies. It supports ontology visualization and reasoning, as well as SPARQL queries. It is a suitable tool for beginners.

owlready2

Owlready2 is a python package for ontology-oriented programming. It has been used to populate ontologies with real data.

Stardog

Stardog is a commercial software for graph data virtualization, knowledge graph exploration, and SPARQL queries.

B The generation of OPLS-AA parameter files in xml format

OpenMM can only process parameter files in xml format. The OPLS-AA parameter file in xml format can be generated by a python script 'processTinkerForceField.py'. The LJ combination rule for OPLS-AA can be performed as described at `http://zarbi.chem.yale.edu/ligpargen/openMM_tutorial.html`.

Bibliography

[1] Waldron, K. J.; Rutherford, J. C.; Ford, D.; Robinson, N. J. *Nature* **2009**, *460*, 823–830.

[2] Andreini, C.; Bertini, I.; Cavallaro, G.; Holliday, G. L.; Thornton, J. M. *J. Biol. Inorg. Chem.* **2008**, *13*, 1205–1218.

[3] Krężel, A.; Hao, Q.; Maret, W. *Arch. Biochem. Biophys.* **2007**, *463*, 188–200.

[4] Maret, W. *Biochemistry* **2004**, *43*, 3301–3309.

[5] Sandstead, H. H. *J. Lab. Clin. Med.* **1994**, *124*, 322–327.

[6] Masters, C. L.; Simms, G.; Weinman, N. A.; Multhaup, G.; McDonald, B. L.; Beyreuther, K. *Proc. Natl. Acad. Sci.* **1985**, *82*, 4245–4249.

[7] Hardy, J.; Selkoe, D. J. *Science* **2002**, *297*, 353–356.

[8] Bush, A. I.; Pettingell, W. H.; Multhaup, G.; Paradis, M.; Vonsattel, J.-P.; Gusella, J. F.; Beyreuther, K.; Masters, C. L.; Tanzi, R. E. *Science* **1994**, *265*, 1464–1467.

[9] Stoltenberg, M.; Bush, A.; Bach, G.; Smidt, K.; Larsen, A.; Rungby, J.; Lund, S.; Doering, P.; Danscher, G. *Neuroscience* **2007**, *150*, 357–369.

[10] Noy, D.; Solomonov, I.; Sinkevich, O.; Arad, T.; Kjaer, K.; Sagi, I. *J. Am. Chem. Soc.* **2008**, *130*, 1376–1383.

[11] Miller, Y.; Ma, B.; Nussinov, R. *Proc. Natl. Acad. Sci.* **2010**, *107*, 9490–9495.

[12] Bell, S. G.; Vallee, B. L. *Chembiochem* **2009**, *10*, 55–62.

[13] Hamer, D. H. *Annu. Rev. Biochem.* **1986**, *55*, 913–951.

[14] Si, M.; Lang, J. *J. Hematol. Oncol.* **2018**, *11*, 1–20.

[15] Li, Y.; Maret, W. *J. Anal. At. Spectrom.* **2008**, *23*, 1055–1062.

[16] Cassandri, M.; Smirnov, A.; Novelli, F.; Pitolli, C.; Agostini, M.; Malewicz, M.; Melino, G.; Raschellà, G. *Cell Death Discov.* **2017**, *3*, 1–12.

[17] Tamames, B.; Sousa, S. F.; Tamames, J.; Fernandes, P. A.; Ramos, M. J. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 466–475.

[18] Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Koksch, B.; Blum, V.; Scheffler, M. *Chem. Eur. J.* **2013**, *19*, 11224–11234.

[19] Zhou, M.; Dong, X.; Baldauf, C.; Chen, H.; Zhou, Y.; Springer, T. A.; Luo, X.; Zhong, C.; Gräter, F.; Ding, J. *Blood* **2011**, *117*, 4623–4631.

[20] Fröhlking, T.; Bernetti, M.; Calonaci, N.; Bussi, G. *J. Chem. Phys.* **2020**, *152*.

[21] Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–520.

[22] Baftizadeh, F.; Biarnes, X.; Pietrucci, F.; Affinito, F.; Laio, A. *J. Am. Chem. Soc.* **2012**, *134*, 3886–3894.

[23] Arkhipov, A.; Shan, Y.; Das, R.; Endres, N. F.; Eastwood, M. P.; Wemmer, D. E.; Kuriyan, J.; Shaw, D. E. *Cell* **2013**, *152*, 557–569.

[24] Schaller, R. R. *IEEE Spectr.* **1997**, *34*, 52–59.

[25] Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[26] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1*, 19–25.

[27] Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D., et al. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

[28] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008**, *29*, 1859–1865.

[29] Camilloni, C.; Pietrucci, F. *Adv. Phys.: X* **2018**, *3*, 1477531.

[30] Vitalini, F.; Mey, A. S. J. S.; Noé, F.; Keller, B. G. *J. Chem. Phys.* **2015**, *142*, 84101.

[31] Schneider, M.; Baldauf, C. *arXiv preprint arXiv:1810.10596* **2018**,

[32] Maksimov, D.; Baldauf, C.; Rossi, M. *Int. J. Quantum Chem.* **2021**, *121*, e26369.

[33] Marianski, M.; Supady, A.; Ingram, T.; Schneider, M.; Baldauf, C. *J. Chem. Theory Comput.* **2016**, *12*, 6157–6168.

[34] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[35] Riniker, S. *J. Chem. Inf. Model.* **2018**, *58*, 565–578.

[36] Li, H.; Ngo, V.; Da Silva, M. C.; Salahub, D. R.; Callahan, K.; Roux, B.; Noskov, S. Y. *J. Phys. Chem. B* **2015**, *119*, 9401–9416.

[37] Rezac, J.; Bím, D.; Gutten, O.; Rulisek, L. *J. Chem. Theory Comput.* **2018**, *14*, 1254–1266.

[38] Kishor, S.; Dhayal, S.; Mathur, M.; Ramaniah, L. M. *Mol. Phys.* **2008**, *106*, 2289–2300.

[39] Van Gunsteren, W. F.; Karplus, M. *Macromolecules* **1982**, *15*, 1528–1544.

[40] Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

[41] Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. *J. Chem. Theory Comput.* **2011**, *7*, 4026–4037.

[42] MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S., et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[43] Vanommeslaeghe, K.; Raman, E. P.; MacKerell Jr, A. D. *J. Chem. Inf. Model.* **2012**, *52*, 3155–3168.

[44] Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

[45] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[46] Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comput. Chem.* **1997**, *18*, 1955–1970.

[47] Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2016**, *12*, 2312–2323.

[48] Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. *J. Chem. Theory Comput.* **2012**, *8*, 2553–2558.

[49] Li, P.; Merz Jr, K. M. *Chem. Rev.* **2017**, *117*, 1564–1686.

[50] Duboué-Dijon, E.; Delcroix, P.; Martinez-Seara, H.; Hladílková, J.; Coufal, P.; Krízek, T.; Jungwirth, P. *J. Phys. Chem. B* **2018**, *122*, 5640–5648.

[51] Kohagen, M.; Lepsik, M.; Jungwirth, P. *J. Phys. Chem. Lett.* **2014**, *5*, 3964–3969.

[52] Duboue-Dijon, E.; Javanainen, M.; Delcroix, P.; Jungwirth, P.; Martinez-Seara, H. *J. Chem. Phys.* **2020**, *153*, 050901.

[53] Akin-Ojo, O.; Song, Y.; Wang, F. *J. Chem. Phys.* **2008**, *129*, 064108.

[54] Maurer, P.; Laio, A.; Hugosson, H. W.; Colombo, M. C.; Rothlisberger, U. *J. Chem. Theory Comput.* **2007**, *3*, 628–639.

[55] Li, P.; Song, L. F.; Merz Jr, K. M. *J. Chem. Theory Comput.* **2015**, *11*, 1645–1657.

[56] Li, P.; Song, L. F.; Merz Jr, K. M. *J. Phys. Chem. B* **2015**, *119*, 883–895.

[57] Riniker, S. *J. Chem. Inf. Model.* **2018**, *58*, 565–578.

[58] Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230–252.

[59] Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

[60] Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

[61] Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.

[62] Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

[63] Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87–110.

[64] Sakharov, D. V.; Lim, C. *J. Am. Chem. Soc.* **2005**, *127*, 4921–4929.

[65] Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell Jr, A. D. *Chem. Rev.* **2016**, *116*, 4983–5013.

[66] Li, Y. L.; Mei, Y.; Zhang, D. W.; Xie, D. Q.; Zhang, J. Z. *J. Phys. Chem. B* **2011**, *115*, 10154–10162.

[67] Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141–6156.

[68] Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730–4737.

[69] Patel, S.; Mackerell Jr, A. D.; Brooks III, C. L. *J. Comput. Chem.* **2004**, *25*, 1504–1514.

[70] Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B.; Friesner, R. A. *J. Chem. Phys.* **1999**, *110*, 741–754.

[71] Yang, Z.-Z.; Wang, J.-J.; Zhao, D.-X. *J. Comput. Chem.* **2014**, *35*, 1690–1706.

[72] Li, P.; Merz, K. M. *Chem. Rev.* **2017**, *117*, 1564–1686.

[73] Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. *J. Phys. Condens. Matter* **2009**, *21*, 333102.

[74] Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

[75] Gresh, N.; Policar, C.; Giessner-Prettre, C. *J. Phys. Chem. A* **2002**, *106*, 5660–5670.

[76] Sakharov, D. V.; Lim, C. *J. Comput. Chem.* **2009**, *30*, 191–202.

[77] Dudev, T.; Lim, C. *Chem. Rev.* **2014**, *114*, 538–556.

[78] Jing, Z.; Liu, C.; Qi, R.; Ren, P. *Proc. Natl. Acad. Sci.* **2018**, *115*, E7495–E7501.

[79] Yu, H.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; MacKerell Jr, A. D.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 774–786.

[80] Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A., et al. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

[81] Lopes, P. E.; Roux, B.; MacKerell, A. D. *Theor. Chem. Acc.* **2009**, *124*, 11–28.

[82] Hu, X.; Lenz-Himmer, M.-O.; Baldauf, C. *Sci. Data* **2022**, *9*, 1–14.

[83] Amin, K. S.; Hu, X.; Salahub, D. R.; Baldauf, C.; Lim, C.; Noskov, S. *J. Chem. Phys.* **2020**, *153*, 144102.

[84] Jensen, F. *Introduction to computational chemistry*; John Wiley & Sons, 2017.

[85] Engel, E.; Dreizler, R. M. *Density Functional Theory*; Springer, 2013.

[86] McWeeny, R. *Nature* **1973**, *243*, 196–198.

[87] Born, M.; Oppenheimer, R. *Ann. Phys.* **1927**, *389*, 457–484.

[88] Hartree, D. R. The wave mechanics of an atom with a non-Coulomb central field. Mathematical Proceedings of the Cambridge Philosophical Society. 1928.

[89] Pauli, W. *Z. Phys.* **1925**, *31*, 765–783.

[90] Sherrill, C. D.; Schaefer III, H. F. *Adv. Quantum Chem.* **1999**, *34*, 143–269.

[91] Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

[92] Jensen, F. *Introduction to computational chemistry.*; John wiley & sons, 2016.

[93] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

[94] Rezac, J.; Hobza, P. *J. Chem. Theory Comput.* **2013**, *9*, 2151–2155.

[95] Riplinger, C.; Neese, F. *J. Chem. Phys.* **2013**, *138*, 034106.

[96] Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

[97] Martin, R. M. *Electronic structure: Basic theory and practical methods*; Cambridge University Press, 2020.

[98] Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

[99] Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048.

[100] Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

[101] Ceperley, D. M.; Alder, B. J. *Phys. Rev. Lett.* **1980**, *45*, 566.

[102] Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.

[103] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.

[104] Grimme, S. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 211–228.

[105] Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. *Sci. Data* **2016**, *3*, 1–13.

[106] DiStasio, R. A.; von Lilienfeld, O. A.; Tkatchenko, A. *Proc. Natl. Acad. Sci.* **2012**, *109*, 14791–14795.

[107] Rossi, M.; Fang, W.; Michaelides, A. *J. Phys. Chem. Lett.* **2015**, *6*, 4233–4238.

[108] Casimir, H. B.; Polder, D. *Phys. Rev.* **1948**, *73*, 360.

[109] Tang, K.; Karplus, M. *Phys. Rev.* **1968**, *171*, 70.

[110] Tang, K. *Phys. Rev.* **1969**, *177*, 108.

[111] Hepburn, J.; Scoles, G.; Penco, R. *Chem. Phys. Lett.* **1975**, *36*, 451–456.

[112] Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

[113] Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

[114] Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

[115] Jurečka, P.; Černỳ, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.

[116] Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2010**, *6*, 1081–1088.

[117] Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

[118] Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.

[119] Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.

[120] Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129–138.

[121] Olasz, A.; Vanommeslaeghe, K.; Krishtal, A.; Veszprémi, T.; Van Alsenoy, C.; Geerlings, P. *J. Chem. Phys.* **2007**, *127*, 224105.

[122] Jurečka, P.; Šponer, J.; Černỳ, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

[123] Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.

[124] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. *Nat. Methods* **2017**, *14*, 71–73.

[125] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[126] A. Kaminski, G.; A. Friesner, R.; Tirado-Rives, J.; L. Jorgensen, W. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

[127] Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

[128] Warshel, A.; Lifson, S. *J. Chem. Phys.* **1970**, *53*, 582–594.

[129] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.

[130] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[131] Jorgenson, W.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

[132] Steiner, D.; Allison, J. R.; Eichenberger, A. P.; van Gunsteren, W. F. *J. Biomol. NMR* **2012**, *53*, 223–246.

[133] Mackerell Jr, A. D.; Feig, M.; Brooks III, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

[134] Kramer, H.; Herschbach, D. *J. Chem. Phys.* **1970**, *53*, 2792–2800.

[135] Halgren, T. A. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.

[136] Lorentz, H. A. *Ann. Phys.* **1881**, *248*, 127–136.

[137] Mohebifar, M.; Johnson, E. R.; Rowley, C. N. *J. Chem. Theory Comput.* **2017**, *13*, 6146–6157.

[138] Liu, Y.-P.; Kim, K.; Berne, B.; Friesner, R. A.; Rick, S. W. *J. Chem. Phys.* **1998**, *108*, 4739–4755.

[139] Ma, B.; Lii, J.-H.; Allinger, N. L. *J. Comput. Chem.* **2000**, *21*, 813–825.

[140] Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694–715.

[141] Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497–1506.

[142] Ren, P.; Wu, C.; Ponder, J. W. *J. Chem. Theory Comput.* **2011**, *7*, 3143–3161.

[143] Anezo, C.; de Vries, A. H.; Höltje, H.-D.; Tieleman, D. P.; Marrink, S.-J. *J. Phys. Chem. B* **2003**, *107*, 9424–9433.

[144] Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902–1921.

[145] Frenkel, D.; Smit, B.; Tobochnik, J.; McKay, S. R.; Christian, W. *Comput. Phys.* **1997**, *11*, 351–354.

[146] Verlet, L. *Phys. Rev.* **1967**, *159*, 98.

[147] Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford University Press, 2017.

[148] Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.

[149] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *J. Comput. Phys.* **1977**, *23*, 327–341.

[150] Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.

[151] Shell, M. S. *Thermodynamics and statistical mechanics: an integrated approach*; Cambridge University Press, 2015.

[152] Berendsen, H. J.; Postma, J. v.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[153] Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

[154] Broyden, C. G. *Math. Comput.* **1967**, *21*, 368–381.

[155] Jorge, N.; Stephen, J. W. *Numerical optimization*; Spinger, 2006.

[156] Hofmann, O. T.; Zojer, E.; Hörmann, L.; Jeindl, A.; Maurer, R. J. *Phys. Chem. Chem. Phys.* **2021**, *23*, 8132–8180.

[157] Clark, D. E.; Westhead, D. R. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337–358.

[158] Cartwright, H. M. *Applications of evolutionary computation in chemistry*; Springer Science & Business Media, 2004; Vol. 110.

[159] Damsbo, M.; Kinnear, B. S.; Hartings, M. R.; Ruhoff, P. T.; Jarrold, M. F.; Ratner, M. A. *Proc. Natl. Acad. Sci.* **2004**, *101*, 7215–7222.

[160] Srinivas, M.; Patnaik, L. M. *Computer* **1994**, *27*, 17–26.

[161] Semenkin, E.; Semenkina, M. Self-configuring genetic algorithm with modified uniform crossover operator. International Conference in Swarm Intelligence. 2012; pp 414–421.

[162] Tsutsui, S.; Yamamura, M.; Higuchi, T. Multi-parent recombination with simplex crossover in real coded genetic algorithms. Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1. 1999; pp 657–664.

[163] Supady, A.; Blum, V.; Baldauf, C. *J. Chem. Inf. Model.* **2015**, *55*, 2338–2348.

[164] Eberhart, R.; Kennedy, J. A new optimizer using particle swarm theory. MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science. 1995; pp 39–43.

[165] Mirjalili, S. *Evolutionary algorithms and neural networks*; Springer, 2019; pp 43–55.

[166] Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607.

[167] Hukushima, K.; Nemoto, K. *J. Phys. Soc. Japan* **1996**, *65*, 1604–1608.

[168] Hansmann, U. H. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

[169] Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.

[170] Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[171] Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

[172] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[173] Stigler, S. M. *Ann. Stat.* **1981**, 465–474.

[174] Ahmed, S. *Technometrics* **2008**, *50*, 238.

[175] Sundberg, R. *Encyclopedia of Environmetrics*; American Cancer Society, 2006.

[176] Marquardt, D. W.; Snee, R. D. *Am. Stat.* **1975**, *29*, 3–20.

[177] Tikhonov, A. N.; Goncharsky, A.; Stepanov, V.; Yagola, A. G. *Numerical methods for the solution of ill-posed problems*; Springer Science & Business Media, 2013; Vol. 328.

[178] Piegorsch, W. W. *Statistical data analytics: Foundations for data mining, informatics, and knowledge discovery*; John Wiley & Sons, 2015.

[179] Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *12*, 55–67.

[180] Tibshirani, R. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288.

[181] Clarke, B.; Fokoue, E.; Zhang, H. H. *Principles and theory for data mining and machine learning*; Springer Science & Business Media, 2009.

[182] Ranstam, J.; Cook, J. A. *Br. J Surg.* **2018**, *105*, 1348–1348.

[183] Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. W.; Silva Santos, D.; Bourne, P. E., et al. *Sci. Data* **2016**,

[184] Berendt, B.; Hotho, A.; Stumme, G. Towards semantic web mining. International semantic web conference. 2002; pp 264–278.

[185] Álvarez-Moreno, M.; de Graaf, C.; Lopez, N.; Maseras, F.; Poblet, J. M.; Bo, C. *J. Chem. Inf. Model.* **2015**, *55*, 95–103.

[186] Řezáč, J.; Jurečka, P.; Riley, K. E.; Černỳ, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J., et al. *Collect. Czechoslov. Chem. Commun.* **2008**, *73*, 1261–1270.

[187] Lenz, M.-O. Towards efficient novel materials discovery: Acceleration of high-throughput calculations and semantic management of big data using ontologies. Ph.D. thesis, Humboldt-Universität Berlin, 2022.

[188] Ghiringhelli, L. M.; Carbogno, C.; Levchenko, S.; Mohamed, F.; Huhs, G.; Lüders, M.; Oliveira, M.; Scheffler, M. *Npj Comput. Mater.* **2017**, *3*, 1–9.

[189] Berners-Lee, T.; Hendler, J.; Lassila, O. *Sci. Am.* **2001**, *284*, 34–43.

[190] Berners-Lee, T. Semantic Web. W3C Web site. `https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html`.

[191] Pfaff, M. Ontology-based semantic data integration in the domain of IT benchmarking. Ph.D. thesis, Technische Universität München, 2018.

[192] Farrell, J.; Lausen, H. *W3C recommendation* **2007**, *28*.

[193] McGuinness, D.; van Harmelen, F. OWL Web Ontology Language Overview. `http://www.w3.org/TR/owl-features/`, 2004.

[194] Gruber, T. R. *Knowledge acquisition* **1993**, *5*, 199–220.

[195] Studer, R.; Benjamins, V. R.; Fensel, D. *Data Knowl. Eng.* **1998**, *25*, 161–197.

[196] Noy, N. F.; McGuinness, D. L., et al. Ontology development 101: A guide to creating your first ontology. 2001.

[197] Piquemal, J.-P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*, 054511.

[198] Zhang, A.; Yu, H.; Liu, C.; Song, C. *Nat. Commun.* **2020**, *11*, 1–10.

[199] Smith, J. S.; Isayev, O.; Roitberg, A. E. *Chem. Sci.* **2017**, *8*, 3192–3203.

[200] Ngo, V.; da Silva, M. C.; Kubillus, M.; Li, H.; Roux, B.; Elstner, M.; Cui, Q.; Salahub, D. R.; Noskov, S. Y. *J. Chem. Theory Comput.* **2015**, *11*, 4992–5001.

[201] Dudev, T.; Lin, Y.-l.; Dudev, M.; Lim, C. *J. Am. Chem. Soc.* **2003**, *125*, 3168–3180.

[202] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., et al. *J. Comput. Chem.* **2010**, *31*, 671–690.

[203] Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. *Nucleic Acids Res.* **2017**, *45*, W331–W336.

[204] Wang, L.-P.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

[205] Kumar, A.; Yoluk, O.; MacKerell Jr, A. D. *J. Comput. Chem.* **2020**, *41*, 958–970.