JCTC Journal of Chemical Theory and Computation

Open Access
This article is licensed under CC-BY 4.0

pubs.acs.org/JCTC Article

System-Specific Parameter Optimization for Nonpolarizable and Polarizable Force Fields

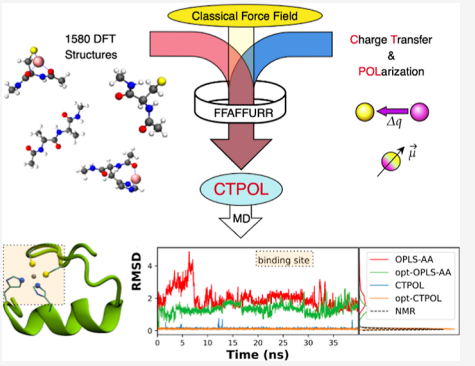
Xiaojuan Hu,* Kazi S. Amin,* Markus Schneider, Carmay Lim, Dennis Salahub,* and Carsten Baldauf*

Cite This: J. Chem. Theory Comput. 2024, 20, 1448−1464

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 (CT) and polarization (POL) effects. In this work, we introduce our implementation of the CTPOL model in OpenMM, which extends the classical additive FF formula by adding CT and POL. Furthermore, we present an open-source parametrization tool, called FFAFFURR, that enables the (system-specific) parametrization 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.



1. INTRODUCTION

Metal ions are essential in biological systems and are involved in physiological functions ranging from maintaining protein structure and 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 ion 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. Multiple fingers can combine together to carry out many complex functions, such as regulating DNA/RNA transcription,[10,11] protein folding and assembly, lipid binding, zinc sensing,[12] and even protein recognition.[13] 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 timescales of milliseconds.[14−16] The majority of current MD studies employ classical force fields (FFs) such as OPLS-AA,[17] AMBER,[18] CHARMM,[19] and GROMOS.[20] It is a challenge for classical force-field models to describe metal−protein interactions due to the strong local electrostatic field and induction effect,[21−26] for example, computer simulation of zinc-containing proteins has been a long-standing challenge that appears hard to tackle without an explicit treatment of charge transfer (CT) or polarization (POL).

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[27] were used to fit parameters to reproduce *ab initio* forces. Empirical continuum correction (ECC)[28−30] force fields scale the charges to implicitly take electronic POL into account. Several works[31,32] 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 parametrization

Received: October 16, 2023
Revised: December 4, 2023
Accepted: December 5, 2023
Published: January 27, 2024

Downloaded via FRITZ HABER INST DER MPI on March 1, 2024 at 11:13:28 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications
© 2024 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

tools, for example, CHARMM general force field (CGenFF),[33] LigParGen,[34] and Antechamber.[35,36] 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.[37] However, the accuracy of assigning approximate parameters to a specific system is limited, and parameters already present in a given FF may also have to be optimized. FFparam[38] and ForceBalance[39] enable the tuning of existing FF parameters. All these parametrization tools share a common assumption of transferability, which assumes that 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 POL effects.[40] LJ parameters are also sensitive to the local environment[41,42] and long-range electrodynamic screening.[43] In this regard, a fundamentally different approach to derive environment-specific or molecule-specific parameters is proposed in refs 44−46. 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 into the model. Including POL effects is a significant step to improve force fields.[47,48] 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 POL plays an important role in the simulation of ion channels,[49] enzymatic catalysis,[50] protein−ligand binding affinity,[51] and dynamic properties of proteins.[52]

At present, there are three main groups of polarizable force fields, fluctuating charge, induced point-dipoles, and Drude oscillator models.[53] The fluctuating charge models simulate POL effects by allowing the charge to flow through the molecule until the electronegativities of atoms become equalized, while keeping the total charge unchanged.[54] One drawback of the fluctuating charge model is that it fails to capture out-of-plane POL of planar or linear chemical groups. The fluctuating charge formula can also be used in conjunction with the Drude oscillator model as a complementary approach to account for CT.[55] A notable model is SIBFA (Sum of Interactions Between Fragments *Ab Initio* Computed).[56]

The induced point-dipole models describe POL energy as the interaction between static point charges and induced dipole moments. Notable induced point-dipole models include OPLS/PFF,[57] AMBER ff02,[58] and AMOEBA.[59,60] 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,[21,61,62] it may be limited when the CT between the cations and coordinating ligand atoms is significant, for example, Cys^- coordinated to metal ions.[63] Ngo et al.[64] and Dudev et al.[65] 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 CT is critical for the development of next-generation polarizable FFs.

The CTPOL[66,67] model incorporates CT and POL effects into classical force fields. The inclusion of CT reduces the amount of partial charge on cations and cation-coordinating atoms. Thus, their charge/dipole−charge interactions are weakened. The local POL 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[23] has shown how QM data[68,69] drives the parameter development of the Drude and CTPOL models. However, most parametrization tools focus on classical force-field models. FFparam[38] provides the parametrization of the Drude model; a CTPOL parametrization tool is not yet available.

In this work, we fill this gap by (i) implementing the CTPOL model in OpenMM[70] and sharing this code[71] and (ii) publishing the Framework For Adjusting force fields Using Regularized Regression (in short FFAFFURR) an open-source tool, which facilitates the parametrization of OPLS-AA and CTPOL models for a specific system, e.g., a peptide system or a peptide−cation system. A major advantage of FFAFFURR is the rapid construction of FFs for troublesome metal centers in metalloproteins. In this work, the new parameters obtained from FFAFFURR were validated by the comparison of FF energies and QM energies in isolation and by assessing the stability of condensed-phase MD simulations 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 for parametrization 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 the 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.

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

where K_{ij}^{r}, K^{θ}, V_n, and $V_{2\mathrm{imp}}$ are force constants, r_{ij}^{0} and θ^{0} are the reference bond length and bond angle, r_{ij}, θ and ϕ are

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

current bond length, bond angle and dihedral angle, respectively, n is the periodicity, φ^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). The index ij refers to a couple of two bonded atoms.

Classical force-field simulations were performed using OpenMM7, a high-performance toolkit for molecular simulations.[70]

2.2. CTPOL Model. The CTPOL[66,67] model introduces charge transfer and POL effects into classical force fields. Instead of a fixed-charge model, CTPOL takes the CT from a ligand atom L (O, S, and N) to a metal cation into account. The amount of transferred charge, Δq_{L-Me}, is assumed to depend linearly on the interatomic distance, r_{L-Me}

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

where a_L and b_L are parameters to be determined that are specific for pairs of ligand L (O, S, N) and a metal cation. The parameters a_L and b_L are of opposite sign, so that the magnitude of CT decreases with distance. The distance at which Δq_{L-Me} becomes 0 is

$$r^0_{L-Me} = -\frac{b_L}{a_L} \tag{4}$$

Beyond this distance, we assume CT to be 0. This approximates real-life CT, which is generally 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

$$q_L = q_L^0 + \Delta q_{L-Me} \tag{5}$$

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

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

$$E_r^{pol} = -\frac{1}{2}\sum_i \boldsymbol{\mu}_i \times \boldsymbol{E}_i^0 \tag{6}$$

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 we optimized here is the atomic polarizability

$$\boldsymbol{\mu}_i = \alpha_i \boldsymbol{E}_i \tag{7}$$

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 in OpenMM via a Python script, which can be found at https://github.com/XiaojuanHu/CTPOL_MD.[71] 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 parametrization protocol on dipeptide and dipeptide−cation systems, we created a quantum chemistry data set. The data set consists of six models: (1) $AcAla_2NMe$ (231 conformers); (2) $AcAla_2NMe + Na^+$ (327 conformers); (3) deprotonated cysteine: $AcCys^-NMe$ (77 conformers), which often acts as an interaction center in metalloproteins; (4) $AcCys^-NMe + Zn^{2+}$ (261 conformers); (5) $AcCys_2^-NMe + Zn^{2+}$ (475 conformers), and (6) $AcHisDNMe + Zn^{2+}$ (209 conformers). 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.[72]

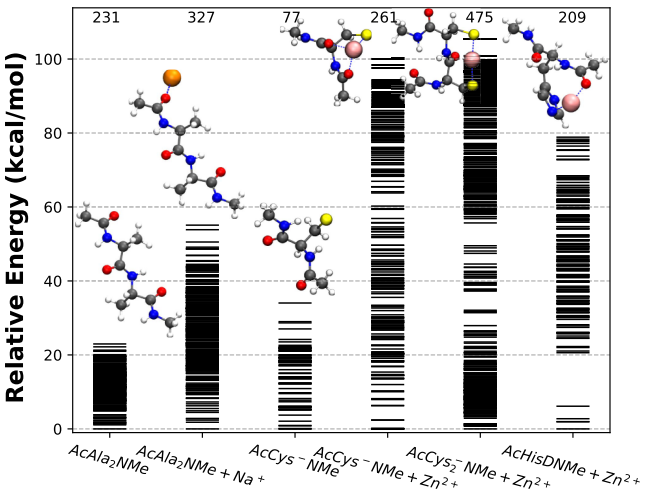


Figure 1. Structures and energy hierarchies of reference data in this study. The numbers of conformers are stated at the top of the individual columns.

All DFT calculations in this work were performed with the numeric atom-centered basis set all-electron code FHI-aims.[73−75] The PBE[76] generalized-gradient exchange−correlation functional augmented by the correction of van der Waals interactions using the Tkatchenko−Scheffler formalism[77] (PBE + vdW^{TS}) was employed. The choice of functional has been validated in previous articles.[68,78] For each conformation, several types of partial charges were provided. Hirshfeld charges[79] are derived based on the Hirshfeld partitioning scheme.[79,80] ESP charges[79,81] are derived by fitting partial charges to reproduce the electrostatic potential. Restrained electrostatic potential (RESP) charges[82] are extracted by a two-stage RESP fitting procedure[82] within the antechamber suite of the AmberTools package.[18] The electrostatic potential was evaluated on a set of grids in a fixed spatial region located in a cubic space around the molecule. The five 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.[83] and Schneider and Baldauf.[25] First, a global conformational search was performed with the basin-hopping approach[84,85] at the force-field level (OPLS-AA).[86] The scan program of the TINKER molecular modeling package[87,88] was employed to perform the basin-hopping search strategy. An energy threshold of 100 kcal/mol for local minima and a convergence

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

criterion for local geometry optimizations of 0.0001 kcal/mol were used. All obtained conformers were relaxed at the PBE + vdW^{TS} level with the *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.[89] 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[73−75] at the PBE + vdW^{TS} level with *tight* basis set. Final conformers were obtained after clustering. Both conformational search protocols have been well validated.[83,89]

2.4. Parameter Optimization. Optimization methods used in this work include LASSO (least absolute shrinkage and selection operator)[90] regression, Ridge regression,[91] and particle swarm optimization (PSO).[92,93] 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[94] library. If the parameters cannot be obtained by solving a set of linear equations, e.g., the CT parameters a_L, PSO was employed. 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.[95]

2.5. FFAFFURR Framework. Force-field parametrization is an optimization problem with three challenging aspects:[96] regarding item 1, the parameters of every energy term in a force field have to be optimized since energy terms and parameters are interdependent and only adjusting a subset may cause inconsistency. Items 1 and 3, training and validation, heavily rely on high-quality data. We use DFT data for comparing potential energies and further validate by MD simulation.

1. The optimization problem has to be defined, which consists of the objective of the optimization and, following this, the selection of training data as well as force-field parameters to adjust.
2. In order to perform the force-field parametrization, a preferably automated procedure has to be implemented. The framework and algorithms used in FFAFFURR are explained in this paper.
3. The obtained set of force-field parameters has to be validated against other data than the training data.

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 useable by a molecular simulation package. FFAFFURR acts as a "wrapper" between the molecular mechanics package openMM[70] and the *ab initio* molecular simulation package FHI-aims.[73−75] 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.[89] 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.[97]

2.5.1. Bond and Angle Parametrization. K_{ij}^{r}, K^{θ}, r_{ij}^{0}, and θ^{0} are empirical parameters of bond-stretching and angle-bending terms. The "spring" parameters K_{ij}^{r} and K^{θ} are unaltered in FFAFFURR. The focus simply lies on the "torsional" and "nonbonded" 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^{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 Parametrization. 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 similar to those fitted to the experimental data.[98] 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.[38,99,100] 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 "torsion contribution" from QM $\tilde{E}_{torsions}^{QM}$ is calculated as

$$\tilde{E}_{torsions}^{QM} = E_{total}^{QM} - E_{nonbonded}^{FF} - E_{bond}^{FF} - E_{angle}^{FF} \quad (8)$$

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

2.5.3. Electrostatic Parametrization. A key difference between FF parameter sets is the origin of the atomic partial charges. Deriving charges from QM data is widely used. The workflow of FFAFFURR tested three choices of partial charges: Hirshfeld,[79,80] ESP,[79,81] and RESP[82] 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 Parametrization. Pair-specific 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.[21] FFAFFURR employs pairwise 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.[101−103] 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

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

$$E_{vdw} = \sum_{i<j} \varepsilon_{ij} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{6} \right] f_{ij} \quad (9)$$

where R_{ij}^{min} refers to the atomic distance where the vdW potential is at its minimum. With the definition of the free and effective atomic volume V^{free} and V^{eff}, 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

$$R_{eff}^{0} = \left(\frac{V^{eff}}{V^{free}} \right)^{1/3} R_{free}^{0} \quad (10)$$

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.[104] Pairwise σ_{ij} can be calculated as

$$\sigma_{ij} = 2^{-1/6} R_{ij}^{min} \quad (11)$$

The ε_{ij} parameter from eq 9 can be tuned by fitting FF LJ energies to reproduce QM vdW energies by LASSO or Ridge regression.

2.5.5. Deriving CT 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 1 or 2. Therefore, we added a correction factor for CN in eq 3

$$\Delta q_{L-Me} = \frac{1}{CN^{k}} (a_L r_{Me-L} + b_L) \quad (12)$$

k, a_L, and r^{cutoff} can be adjusted by PSO. b_L can be calculated with the assumption that CT is zero at the cutoff distance. The target objective of fitting can be the QM potential energy, the electrostatic potential derived from electron densities, or, as we do in this work, the QM interaction energy. The correction factor CN is determined by scanning the *ab initio* data set and is thus a constant throughout the MD simulation for a particular atom.

2.5.6. POL Energy. To get the value of atomic polarizability α_i in eq 7, we use the definition of effective polarizability of an atom in a molecule, where the free-atom polarizability is scaled according to its close environment with a partitioning

$$\alpha_{eff} = \left(\frac{V^{eff}}{V^{free}} \right) \alpha_{free}^{0} \quad (13)$$

where V^{eff} and V^{free} are the same as in eq 10, and α_{free}^{0} 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 slightly adjusting 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 CT and POL, we reduce the dimensionality of the energy function and therewith the ability to correctly/fully represent the PES. Consequently, a force field, describing, e.g., such a cation−protein system, cannot fully reproduce a DFT PES. Hence, it appears advisable to put focus on the accuracy of distinct areas of the PES. RMSD between two surfaces is a common fitting criterion, 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

$$\text{wRMSD} = [\sum_{i=1}^{N} w_i (E_i^{FF} - \Delta E_i^{QM})^2]^{1/2} \quad (14)$$

where RMSD is modified by including a Boltzmann-type factor

$$w_i = A \exp \left[\frac{-E_i^{QM}}{RT} \right] \quad (15)$$

where A is the normalization constant (so that $\sum w_i = 1$) and RT is a temperature factor that has no physical meaning in the context of this application, but affects the flatness of the distribution. Our previous work[23] has shown how Boltzmann-type weighted RMSD with an appropriate choice of RT can be utilized as the 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 parametrization, 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

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\Delta E_i^{FF} - \Delta E_i^{QM} + c| \quad (16)$$

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

$$\text{ME} = \max_{i \in N} |\Delta E_i^{FF} - \Delta E_i^{QM} + c| \quad (17)$$

2.6.2. Molecular Dynamics Simulations. We performed MD simulations of the NMR structure 1ZNF[105] with different parameter sets to evaluate the performance of FFAFFURR. All MD simulations were performed using OpenMM7.[70] 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 the 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 method.[106] 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, CT and induced dipoles were updated every 10 steps. Covalent bonds and water angles were constrained.

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

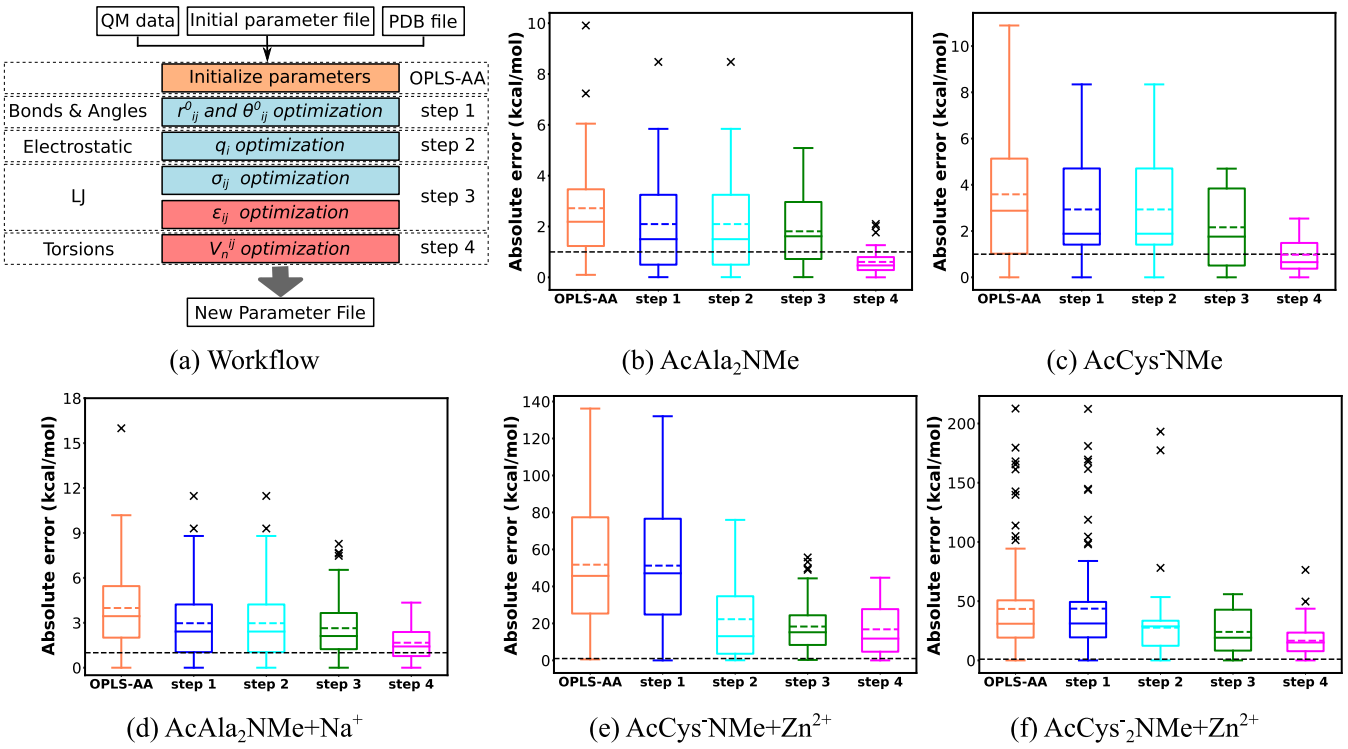


Figure 2. (a) Workflow of the parametrization of OPLS-AA in four major steps. Different colors represent different fitting methods. Parameters in blue boxes are derived from DFT calculation, and parameters in red boxes are tuned by LASSO or Ridge regression. (b–f) Box plots of absolute errors of OPLS-AA parametrization major steps (OPLS-AA, step 1, step 2, step 3, and step 4) for the test set of (b) $AcAla_2NMe$, (c) $AcCys^-NMe$, (d) $AcAla_2NMe + Na^+$, (e) $AcCys^-NMe + Zn^{2+}$, and (f) $AcCys_2^-NMe + Zn^{2+}$. 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 dashed 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 dashed line indicates the chemical accuracy, which is 1 kcal/mol. Note the large differences in scales in subfigures (b–f).

3. RESULTS AND DISCUSSION

To assess the performance of FFAFFURR and describe which protocol to use to create a 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 examining the structural stability of the zinc-finger motif in MD simulations.

3.1. OPLS-AA Parametrization. Although studies have shown that it is difficult to implicitly incorporate the POL effects into classical FFs,[23,107] fine-tuning parameters of fixed-charge models to describe cation–protein systems is still attractive due to its low computational cost and easier parametrization. Here, we tested the performance of the fixed-charge model OPLS-AA parametrized using 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 reference models since the CT and POL effects caused by Na^+ are less than that of Zn^{2+}. On the contrary, Cys^- is one of the ligands that interact with Zn^{2+} in proteins, and CT between Cys^- and Zn^{2+} is significant. For each system, 80% of the conformers were randomly selected as the training set, and the remaining 20% were used as the test set.

We first demonstrate the functionality of FFAFFURR on the example of OPLS-AA parametrization. The key steps of OPLS-AA parametrization are briefly described in Figure 2a. We showed the ability to reproduce PES by optimizing the 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 2a, the parameters in blue boxes are derived from DFT calculations, and the parameters in red boxes are fitted by LASSO or Ridge regression, as described in Sections 2.4 and 2.5. Here, we only tested RESP partial charges, the LASSO method in deriving ε_{ij}, and Ridge regression in deriving V_n^{ij}. The parameters derived from DFT calculations are tuned first because they are considered fixed with respect to changes of the other parameters; then, different tuning orders of the parameters for the other energy terms in the FF formula are tested to choose the order that gives the smallest errors between DFT and FF energies. The final order of the protocol is shown in Figure 2a.

Figure 2b–f shows the comparison of FF energies with optimized parameters after each step in Figure 2a. 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 2e,f indicates 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 CT to some extent. In the case of $AcAla_2NMe$ and $AcCys^-NMe$, the MAEs were improved from 2.72 and 3.59 kcal/mol to 0.61 and 0.98 kcal/mol, respectively, which is well within the chemical accuracy of 1 kcal/mol. In the case of $AcAla_2NMe + Na^+$, the MAE was improved from 3.99 to 1.67 kcal/mol. Although the optimized MAE is above the chemical accuracy, the maximum error is significantly

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448–1464

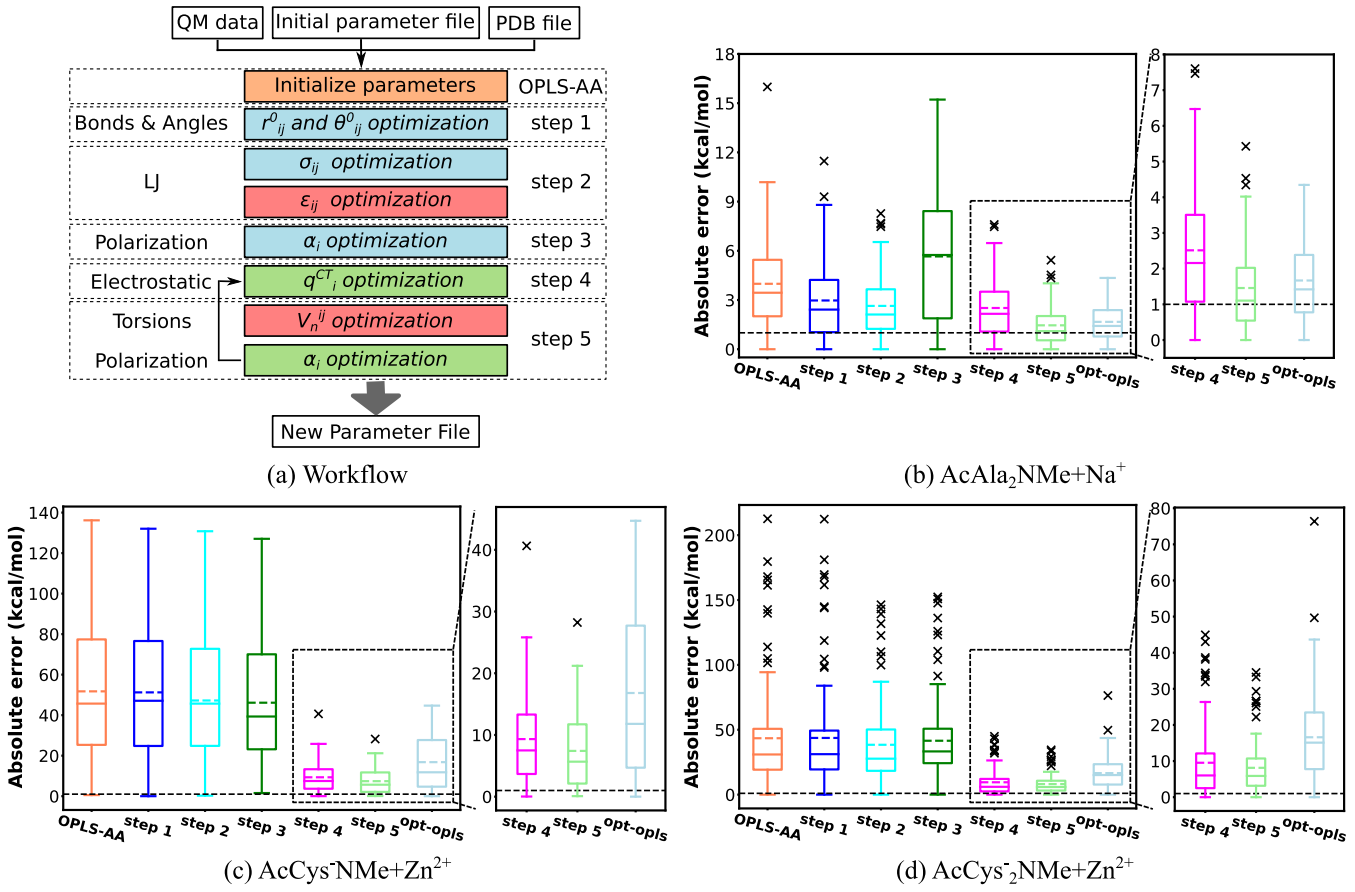


Figure 3. (a) Workflow of full CTPOL parametrization in five major steps. Different colors represent different fitting methods. Parameters in blue boxes are derived from DFT calculation, parameters in red boxes are tuned by LASSO or Ridge regression, and parameters in green boxes are tuned by PSO. (b–d) Box plots of absolute errors of CTPOL parametrization major steps (OPLS-AA, step 1, step 2, step 3, step 4, step 5) and OPLS-AA with full optimized parameters (opt-opls) for the test set of (b) $AcAla_2NMe + Na^+$, (c) $AcCys^-NMe + Zn^{2+}$, and (d) $AcCys_2^-NMe + Zn^{2+}$. 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 dashed 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 dashed line indicates the chemical accuracy, which is 1 kcal/mol.

reduced. However, in the cases of $AcCys^-NMe + Zn^{2+}$ and $AcCys_2^-NMe + Zn^{2+}$, the MAEs were improved from 51.75 and 43.47 kcal/mol to 16.8 and 16.59 kcal/mol, respectively. Although these are by numbers great improvements, the MAEs are much higher than for the other systems. Calculations based on parameters of such quality have no predictive power. This confirms the necessity of explicitly including CT and POL effects to describe divalent ion–dipeptide systems. We note that for dipeptides and dipeptides with monovalent cation systems, optimization of torsional parameters has the greatest impact on the improvement of the accuracy. Previous studies by some of us[78,108] have shown that mono- and divalent cations strongly modify the preferences of torsion angles. While for dipeptides with divalent cations, apparently, the adjustment and treatment of charge interactions play the most important role. This further confirms that the capture of CT and POL is crucial for the accurate description of systems with divalent cations. We also note that the maximum errors are greatly reduced after the parametrization of LJ interactions of the five systems.

3.2. CTPOL Parametrization. The CTPOL model introduces both local POL and CT 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 parametrization workflow are depicted in Figure 3a. Following the methodology of OPLS-AA optimization, parameters unaffected by others are adjusted first; then, different orders are tested to choose the order that gives the smallest errors between DFT and FF energies. Charges are taken from OPLS-AA in step 1 to step 3. In step 4, CT was introduced. As already mentioned, the parameters in blue boxes are derived from DFT calculations, and the parameters in red boxes are fitted by LASSO or Ridge regression. 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 the *ab initio* method. In step 5, we tried to slightly tune α_i by PSO. An additional round of parametrization from step 4 to step 5 can be performed to better optimize the FF parameters.

Absolute errors of each step in Figure 3a are illustrated in Figure 3b–d. Absolute errors of optimized OPLS-AA (opt-opls) are also shown in Figure 3 to compare the performance of FFAFFURR on OPLS-AA and CTPOL models. As shown in Figure 3, the introduction of POL effects in step 3 did not improve the accuracy much, and the errors of the $AcAla_2NMe + Na^+$ system even increased. This may be due to the fact that classical force fields already take some account of POL effects, since the charges come from fitting to reproduce quantum mechanical or experimental electrostatic field distributions.[67]

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448–1464

Including CT from ligand atoms to the cation reduces atomic charges, therefore compensating for the electrostatic potential. Not surprisingly, errors are significantly reduced after including CT, as displayed in Figure 3. After the parametrization, the MAEs of $AcAla_2NMe + Na^+$, $AcCys^-NMe + Zn^{2+}$ and $AcCys_2^-NMe + Zn^{2+}$ reached 1.45, 7.42, and 8.12 kcal/mol, respectively. In contrast, the MAEs of the optimized OPLS-AA are 1.67, 16.8, and 16.59 kcal/mol, respectively. Apparently, the inclusion of CT and POL effects better describes systems involving cations than classical force fields, especially for systems with divalent cations.

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 the CT parameters. In Figure 4, the $AcCys^-NMe + Zn^{2+}$ 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. For all RT, the weight decreases as the relative energy increases, but increasing RT decreases the weights on low-energy conformations. Figure 4 shows the difference in mean absolute errors between unweighted fitting and weighted fitting with RT = 16 kcal/mol. 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 substantially in the low-energy region, while high-energy regions do not get worse.

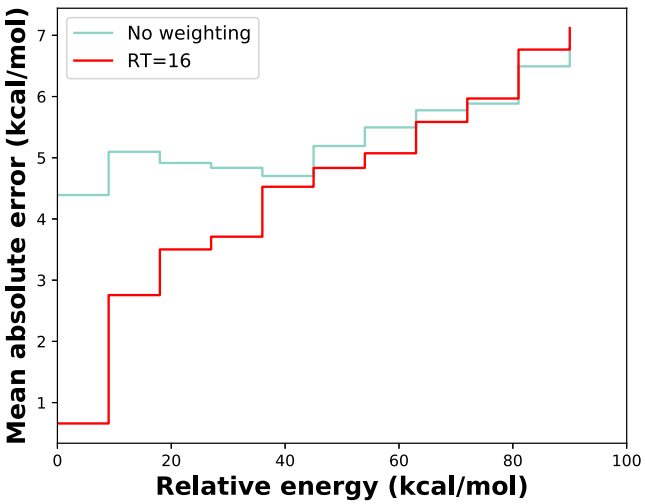


Figure 4. Absolute errors of optimized FF energies with respect to QM energies for the $AcCys^-NMe + Zn^{2+}$ system by weighted fitting (with RT = 16 kcal/mol) and by unweighted fitting. The height of a plateau represents the mean absolute error for conformers whose relative energies are smaller than the right node of the plateau.

3.3. Validation with Molecular Dynamics Simulations.

The 1ZNF PDB structure[105] 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 Zn^{2+} 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 parametrization workflow. One potential application of FFAFFURR to this system is to optimize selected parameters for the interaction center (Figure 5 bottom-left), since that is the region of most complexity. It is important to note that the 1ZNF structure itself is based on a model, where NOE data from NMR is used as restraints for minimization using an AMBER force field to actually obtain the final structure.[105] However, the authors of the paper for 1ZNF also state that the Zn−S and Zn−N distances are restrained within 0.05 Å of 2.30 and 2.00 Å, respectively, and that this agrees well with other X-ray structures of zinc-finger complexes. They also state that the angles were restrained to 109 ± 10°, also in agreement with X-ray structures. This makes this structure ideal as a reference for binding-site metrics such as distances and angles.

In this paper, we used an approach similar to that of Li and Merz,[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 parametrization, 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, as shown in Figure 6, almost

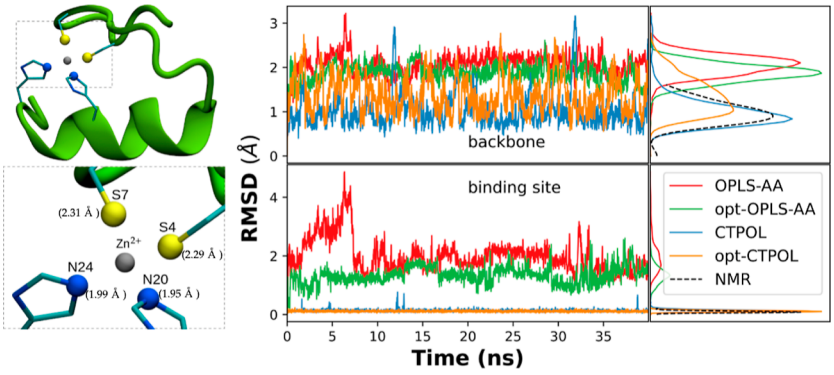


Figure 5. Top left: The protein structure of 1ZNF, with the backbone represented by a ribbon, and the Zn^{2+}-binding site shown explicitly. Bottom left: Zoom in of the site, with distances of the coordinating atoms relative to Zn^{2+}. The sulfurs are from Cys4 and Cys7, while the nitrogens are the NE2 nitrogens of His20 and His24. *Right*: RMSDs of MD trajectories from the NMR structure of 1ZNF (model 1), calculated for different parameter sets, for the backbone (top) and interaction site (bottom). 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 the NMR data of 1ZNF with respect to the first model of the PDB.

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

Table 1. Parameter Sets Used for MD Simulation[a]

parameter set	pairwise 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

[a]The determination of LJ parameters from FFAFFURR is described in Section 2.5. Optimized parameters are listed in Tables S2 and S3.

immediately after the start of the simulation. To try and prevent this, we optimized OPLS-AA according to the steps outlined in Figure 2. Specifically, we optimized the pairwise LJ parameters between atoms in HisD and Zn^{2+} using AcHisDNMe + Zn^{2+} as the target data (Figure 1). The parameters that are optimized are listed in Table S2. The LJ parameters between the atoms in Cys and Zn^{2+} are kept untouched since we have not seen strange behaviors between Cys and Zn^{2+}. We performed similar LJ parameter optimization for the opt-CTPOL model as well. In the CTPOL and opt-CTPOL models, CT was introduced for S/N/O/Zn atoms in the binding site, and POL effects between non-hydrogen atoms and Zn^{2+} were added. Oxygen is included as some of the structures in the *ab initio* data set have Zn^{2+} interacting with a peptide oxygen atom.

3.3.1. Backbone Structure and Binding Domain are Better Preserved with CTPOL. We ran three 40 ns long simulations with each of the four models listed in Table 1, for simplicity, we show only one of the respective simulations in, for example, Figure 5. We also used the 37 experimental NMR structures of 1ZNF to compare the structural features between our simulations and NMR observations. Figure 5 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 same first model to see how much variation occurs among those.

It is clear from Figure 5 that both the overall structure and binding domain are in better agreement with the NMR structures when CT 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 5. This is primarily due to the histidines breaking away from the binding with Zn^{2+}, as supported by Figure S2.

The RMSD values of OPLS-AA and opt-OPLS-AA deviate far from the NMR model, particularly the RMSD values 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 the optimization of the pairwise LJ parameters of Zn^{2+} and histidine (opt-OPLS-AA), we observed one of the histidines escaping the binding domain. Figure 6a,b shows snapshots of such conformations after 40 ns. Similar problems with the binding domain stability have been observed in previous studies, where Zn^{2+} escapes from the coordination center in nonpolarizable FF simulations.[107,112]

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 6c,d). This emphasizes that explicitly including CT and POL effects is critical for a proper description of the binding domain and hence the overall structure of zinc-fingers.

3.3.2. LJ Parametrization Makes the CTPOL Model More Robust. To evaluate the effect of optimized pairwise LJ parameters, we compared the CTPOL model without any LJ parametrization (CTPOL) to the CTPOL model with LJ parametrization (opt-CTPOL). From Figure 5, 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 1 ns 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 a similar behavior, as they are taken from a simulation. Figure 7 shows

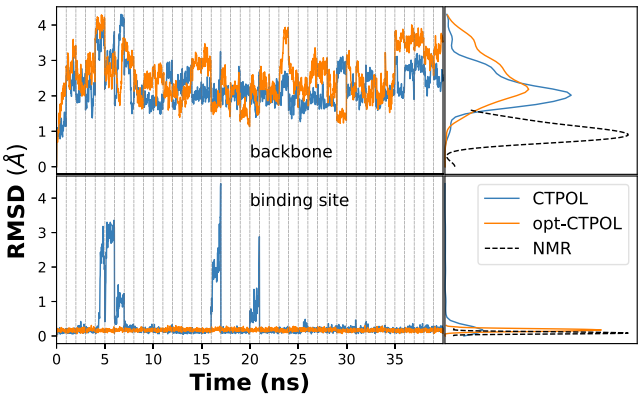


Figure 7. RMSD of CTPOL and opt-CTPOL vs first model of NMR, with 40 trajectories of 1 ns concatenated into one. The dotted lines represent concatenation boundaries of the trajectories.

that the 40 ns trajectory of CTPOL using the NMR structure as the starting point is more or less stable. However, when

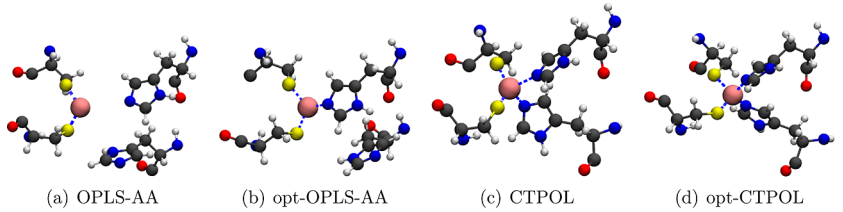


Figure 6. Snapshots showing the conformation of binding site after 40 ns of simulation.

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

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 CTs to Zn^{2+} in CTPOL, as seen in Figure 8. This occurs around the same time as the

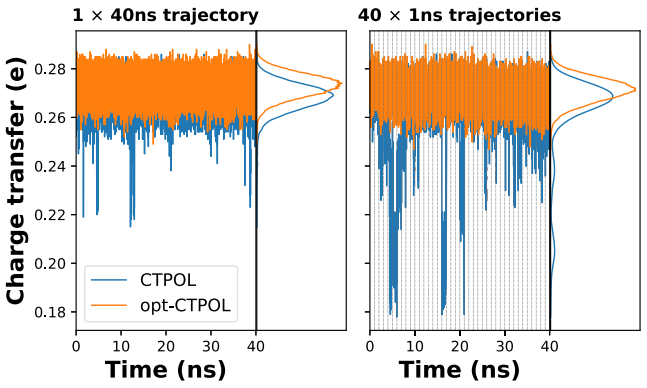


Figure 8. CT as a function of time for (left) a continuous 40 ns trajectory from one stable initial structure and (right) 40 independent 1 ns simulations concatenated together. The dashed vertical lines mark the concatenation boundaries. The 40 × 1 ns simulations were started from different initial conditions randomly chosen from a continuous MD simulation with randomized velocities.

binding domain fluctuations in Figure 7. A closer inspection of the distances between Zn^{2+} and coordinating nitrogens (Figure 9) reveals that these fluctuations are perfectly correlated with

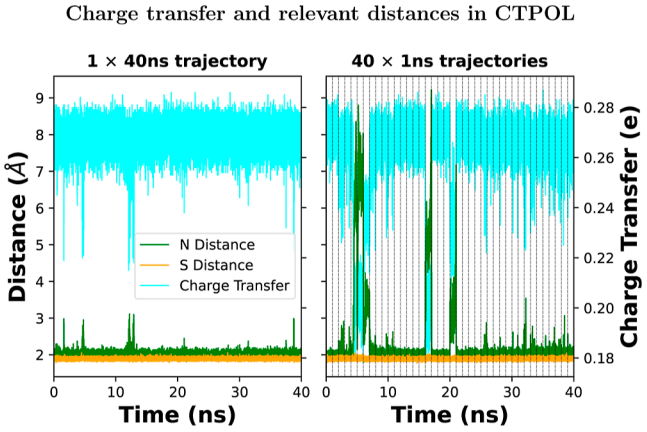


Figure 9. Coordinating nitrogen and sulfur distances (left *y*-axis) and CT (right *y*-axis) vs time for a continuous trajectory (left) and 40 independent concatenated 1 ns trajectories (right). In cyan, we have the CT, in green, the average of the distances of Zn−N20 and Zn−N24, and in yellow the average 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 eight times.

these distances. As the binding site breaks 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 CT contribution of the nitrogens drops to zero, and the only contribution is from the sulfurs, and hence the lower total CT. However, opt-CTPOL appears to have no such fluctuation in either the 40 ns or 40 × 1 ns trajectories.

3.3.3. CT and Relevant Distances in CTPOL. These unfolding events within 1 ns occur about 20% of the time for CTPOL, thus making CTPOL without LJ optimization unreliable.

3.3.4. Opt-CTPOL Shows Improvement with a Caveat to be Addressed in the Future. 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 10 (top). We can see immediately that

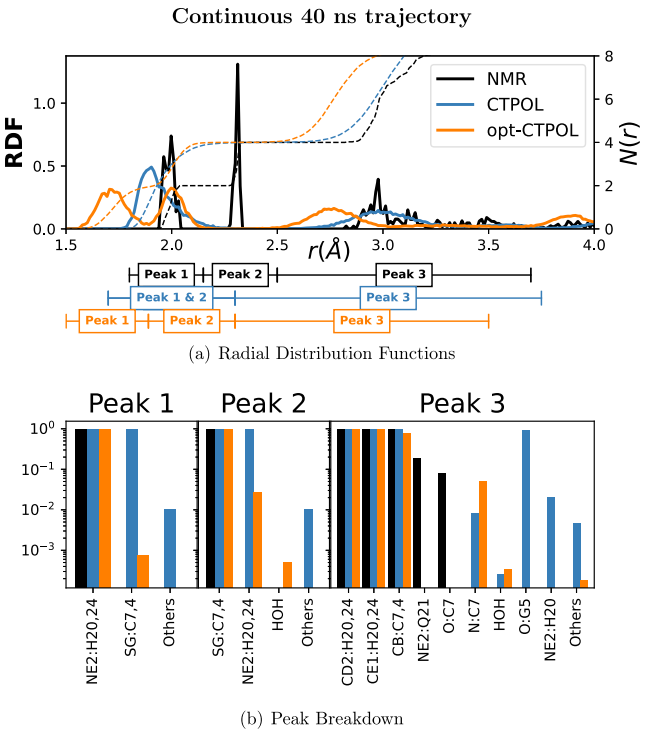


Figure 10. Coordination analyses of continuous 40 ns trajectory. (a) RDF and coordination number ($N(r)$) of all non-hydrogen protein atoms, with the distance ranges of selected peaks. The solid lines are the RDF (left *y* axis), and the dashed lines are the corresponding $N(r)$ (right *y* axis). (b) 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 appears within the peak range.

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. In the NMR models, the first and second peaks at 2.0 and 2.3 Å correspond to Zn^{2+}−N(His) and Zn^{2+}−S(Cys), respectively. In contrast to CTPOL, the opt-CTPOL peaks are distinct, with only a small percentage (<2%) of trajectories showing nitrogens in the second peak dominated by sulfur. These features are also seen in similar analyses of the 40 × 1 ns trajectories (Figure 11). Based on the analyses in this manuscript, we can conclude 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 10 (top)], we determined which atoms comprise each peak and at what fraction of the trajectory these atoms remain in that peak, as shown in Figure 10 (bottom). The first and second peaks in CTPOL appear to be contaminated by other atom types which do not appear in NMR peaks at all. In the 40

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

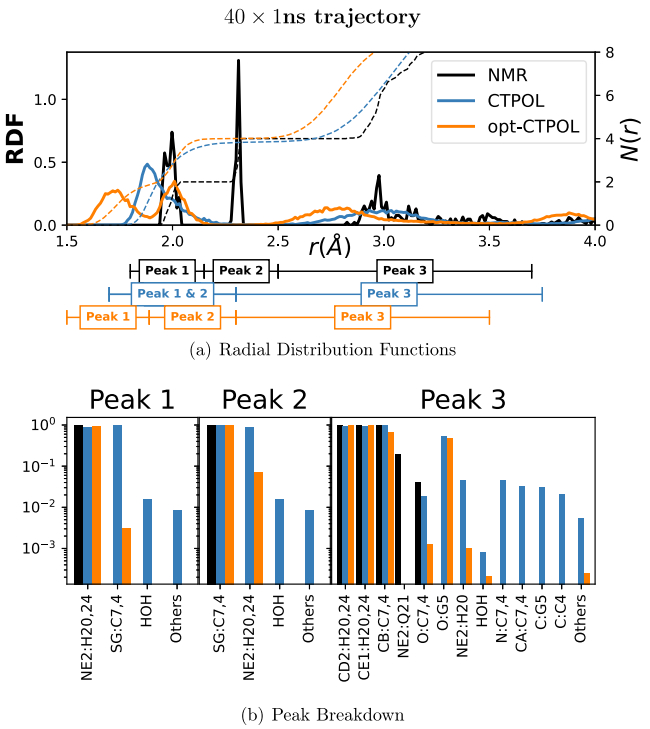


Figure 11. Coordination analyses of 40 × 1 ns trajectory. (a) RDF and coordination number ($N(r)$) of all non-hydrogen protein atoms, with the distance ranges of selected peaks. Solid lines are RDF, and dashed lines are $N(r)$. (b) Composition of each peak, where atoms of the same type and residue are lumped together. The y axis represents the average fraction of conformations in which each of the atoms appears within the peak range.

× 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 11 (bottom)]. The opt-CTPOL model has no other atom types in the first peak and only relatively few others in the second peak not present in NMR.

We should note that the NMR model we used does not contain any explicit water molecules. To determine if water could be present in the binding site, we looked at 15 zinc-finger X-ray crystallography structures from the Protein Data Bank[113] (PDB) Web site (http://www.rcsb.org/pdb/) to find the binding sites which are similar to this one (see Table 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 third 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 first and second peaks, as is the case in the CTPOL model, is uncharacteristic of zn-finger binding sites of similar nature to 1ZNF. The opt-CTPOL model does a better job of keeping water outside these peaks, with only a small fraction of water in the second peak.

3.3.5. Angle and Distance Distributions. 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 12

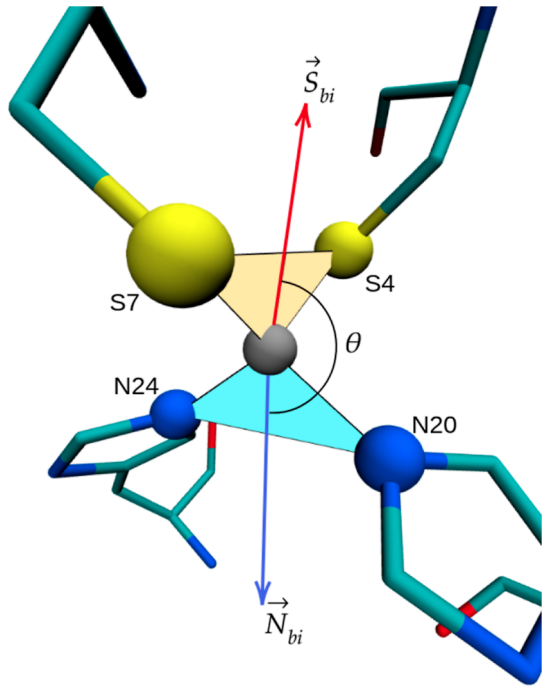


Figure 12. Binding site with Zn^{2+} at the center (gray atom), the sulfurs from Cys7 (S7) and Cys4 (S4), and the NE nitrogens from His20 (N20) and His24 (24). Hydrogens have been removed for clarity. The yellow triangle on top has vertices on Zn, S4, and S7, while the blue triangle at the bottom has vertices on Zn, N20, and N24. The angles between the planes of these triangles are used for plotting the distributions in Figure 14. 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 in Figure 14b. The distributions of some of the distances between the five atoms shown in this figure are shown in Figure 15, while the distributions for some of the angles are shown 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 × 1 ns trajectory where the binding domain destabilizes.

Figure 13 shows the distribution of most of the angles that the coordinating atoms make with Zn^{2+}. Additionally, Figure 14a shows the distributions of angles between the planes shown in Figures 12, and 14b shows the distributions of the angles between the bisectors, also defined in Figure 12. 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 5, 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 10.

Furthermore, we see from Figure 15 that the distances of the opt-CTPOL binding domain are consistently shorter than

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

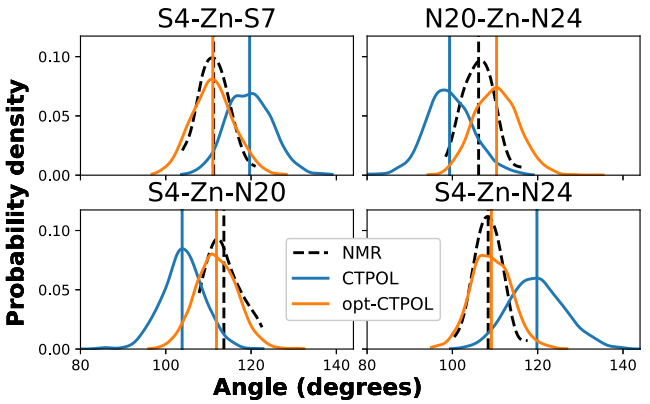
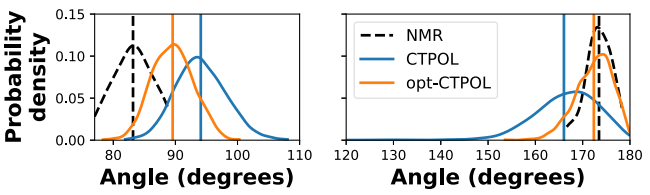


Figure 13. Probability distribution of angles over the continuous 40 ns trajectories of CTPOL (blue) and opt-CTPOL (orange) and over 37 NMR models (black dashed). The corresponding atoms are depicted in Figure 12. The distributions were calculated using kernel density estimation.[110,111] The vertical lines represent the averages of each distribution.



(a) Dihedral distribution (b) Bisector angle distribution

Figure 14. (a) Probability distributions of angles between the S7−Zn−S4 and N24−Zn−N20 planes, as depicted in Figure 12. (b) Angles between S4−Zn−S7 and N20−Zn−N24 bisectors, which are depicted in Figure 12 as $\vec{S}_{bi}$ and $\vec{N}_{bi}$, respectively.

those of the experimental NMR structures. This is in line with the RDF analysis of Figure 10, where we see a 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 opposite, with plenty of overlaps between the two distributions, and thus their first and second peaks in Figure 10 also overlap.

3.3.6. Issues We Observe and Their Possible Origins. The main issue we face is clearly the contraction of distances in the Zn^{2+}-binding site in comparison to NMR data. To investigate the source of this issue, we first analyzed the relevant distances in the *ab initio* data set, as depicted in Figure 16. We see that the Zn−N distances are mostly around the 2 Å mark, and Zn−S distances are mostly around 2.3 Å, as it is indicated by the maxima of the histograms. In both cases, they cover a wide range of energies as well, suggesting that these are preferred distances for various conformations of the dipeptide. However, these are (certainly almost exclusively) minima on the PES.

The larger distances in Figure 16 are due to the coordination of Zn with other atoms (such as carbonyl oxygen). To make a meaningful comparison with the 1ZNF system, we further filtered the QM data down to where the nearest atom to Zn is atom SG or atom NE2, since that is the case in the 1ZNF system. From the filtered data, we computed the average distances and compared them in Table 2, where it is clear that

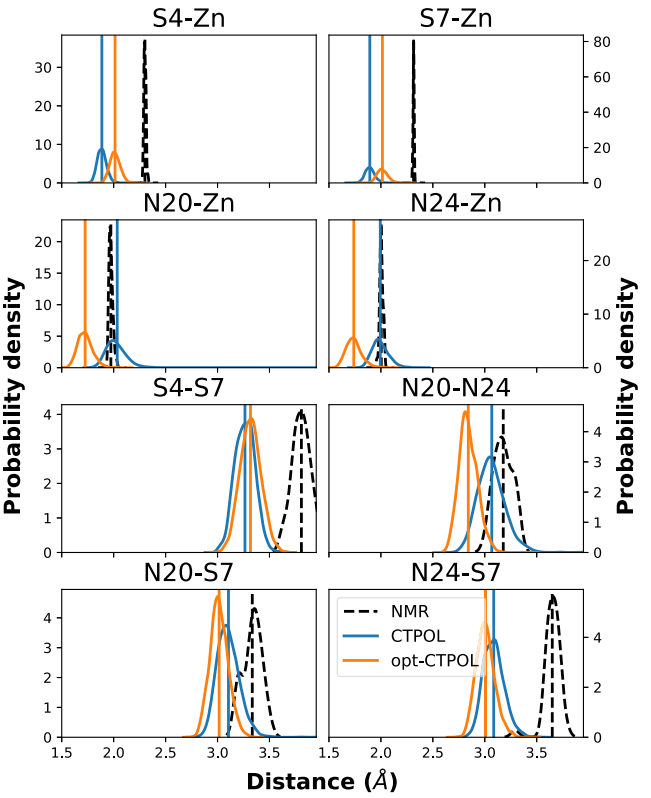


Figure 15. Probability distribution of distances (using kernel density estimation[110,111]) over the entire trajectory (for simulations) and over 37 models (for NMR data).

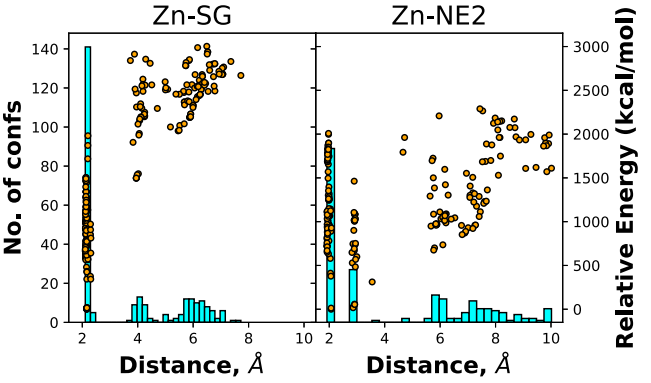


Figure 16. Distribution of *ab initio* distances, with bars representing histograms of the numbers of conformers (left axis) over the distance range and points representing the distance and energy of individual conformers (right axis). The plot on the left takes distances between Zn^{2+} and S(Cys) from $AcCys^{-}NMe + Zn^{2+}$ *ab initio* conformations. The plot on the right takes distances between Zn^{2+} and N(His) from $AcCys^{-}His + Zn^{2+}$ *ab initio* conformations.

Table 2. Average Distance Data

pair	QM	NMR[a]	Others[b]	CTPOL	opt-CTPOL
N−Z	1.95	1.99	2.00	2.01	1.73
S−Z	2.28	2.31	2.30	1.89	2.01

[a]Derived from the data of 1ZNF. [b]Derived from other PDBs (see Table S4 for a full list).

the QM data agree well with the NMR data of 1ZNF as well as the other PDB structures that we used for analyzing water

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

distances. On the other hand, opt-CTPOL systematically gives shorter distances.

Given that the shorter distances are not an artifact that was carried from the gas-phase QM data, in the following we discuss a few likely causes for this behavior of our model:

- We demonstrate in this paper (see Figure 3) that our approach can properly match energy hierarchies. However, the focus on equilibrium geometries may result in insufficient treatment/parametrization of repulsive (off-equilibrium) geometries. The effect may even be amplified due to the Boltzmann-weighting used in the parameter optimization process.
- The LJ-(12,6)-potential is intended to model pairwise repulsive and non-Coulomb attractive interactions among atoms. However, due to the nature of classical force-field formulations, it may also contain aspects of POL, dipole−dipole interactions, and so forth. These aspects may be incorporated in parameters as well as in the functional form (e.g., in the 12−6 form).
- Although we are matching MM energies at the exact same conformations, we do not know if the MM minima themselves correspond to those conformations. It is possible that we are simply matching the MM energies at the QM minima, but the MM minima may be in different states altogether, in this case, those that result in shorter ligand distances.
- None of the *ab initio* references contain both sulfur and nitrogen in the coordination of Zn^{2+}, or even specifically S−N, N−N, and S−S pairwise interactions. We only parametrize Zn^{2+}−ligand LJ interactions in this work as an example, but it is possible that optimizing ligand−ligand interactions will improve the result.

4. CONCLUSIONS AND OUTLOOK

The availability of sufficiently 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. The benefits of such an approach was shown in previous work[23] on the QM-driven parametrization of Drude and CTPOL models for ion−protein interactions in MD simulations.

Since the explicit Drude model includes POL as a degree of freedom subject to forces, it requires shorter time steps and a dual thermostat, and it introduces additional parameters such as the mass of Drude particles and spring constants. CTPOL—as an implicit model—requires the minimization of dipole moments at every time step, but allows for normal time steps. Both models exhibit their own challenges and—in particular our reimplementation of CTPOL—their own room for improvements.

FFAFFURR is developed as a Python tool to facilitate the parametrization of classical and polarizable CTPOL models. In this paper, we chose to parametrize OPLS-AA as an example, although the tool can be adjusted to work with other similar force fields such as CHARMM and AMBER once the code is generalized. It automatically parses QM calculation outputs from FHI-aims and generates parameter files that can be directly processed by the molecular dynamics package OpenMM. FFAFFURR also allows users to choose which energy terms to adjust.

We utilized an extensive data set of model peptide−cation conformers from DFT calculations. Structures cover an extended relative energy range, and all required properties for the parametrization were extracted. Due to the sheer size of the utilized data set—by means of number of conformers and energy range—we think that we have covered sufficient individual structural diversity to properly derive parameters for the FF energy terms. The performance of optimized parameters in each energy term was evaluated by comparing FF energies and QM potential energies.

We showed that the CTPOL model outperforms OPLS-AA in terms of the accuracy of reproducing the QM energy hierarchies 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[105] and comparing the simulation results with NMR models. With the parameters optimized from FFAFFURR, we found that CTPOL much better reproduces the overall structure of the protein, while with OPLS and opt-OPLS, the Zn^{2+} binding site unravels. 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 agree with the NMR models than CTPOL alone.

On the other hand, the optimization of LJ does lead to a shrunken binding domain. While we briefly discussed reasons for this behavior above, we focus here on possible remedies: in summary, FFAFFURR has a wide range of functions to facilitate and perform parameter optimization in peptide systems. It can provide better additive force-field parameters for systems with no cations, as seen from Figure 2b,c. Additionally, FFAFFURR can provide almost all the functions required for the cation−peptide parametrization process, including CTPOL force fields. FFAFFURR helps users remove labor-intensive steps in force-field optimization.

- A possible lack of off-equilibrium geometries in the parametrization could be tested first by omitting the Boltzmann-weighting. This would give more impact to higher-energy structures that are more likely to feature at least partly more compressed structures. Maybe, this already would lead to improved geometries.
- The repulsive part of the LJ potential could be better captured by locally exploring the PES by adding points through dislocating atoms in the cation interaction site. This could either be realized by carefully altering Cartesian coordinates by small values in the ± space directions or through short DFT-based MD simulations. This would also capture whether or not the MM-minima are at those conformations.
- The functional form of the van der Waals potential may be better generalized by alternatives to the LJ form which allow for better tuning of the shape of the repulsive components. For example, the Mie potential is a generalized version of LJ, where the exponents and coefficients of each term can be varied. The Buckingham potential is another alternative where the repulsive term is replaced with an exponential decay with variable amplitude and decay rate. Such adjustments of the van

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

der Waals potential also require a database that contains sufficient off-equilibrium geometries.

- Lastly, the data set could be expanded to contain models with more than one amino acid side chain, in order to capture complex coordination structures, allowing us to parametrize other pairwise interactions such as sulfur and nitrogen and not just Zn^{2+} and ligand atoms.

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 have to be investigated in more detail in a future study.

Further improvements can also be made to the evaluation of FFAFFURR using MD simulations. This paper uses zinc-finger binding site as it is reference, which does provide a good initial test for binding site stability and NMR structure reproduction. However, since one issue may be overbinding, it is worthwhile to expand our experimental reference structures to also include systems in other states, such as unbound metal ion or multimeric systems. In this regard, one useful characteristic of FFAFFURR is that it can be employed to derive parameters for a specific system in order to grasp the specific environment of the system in a classical force field. 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.[69] If the user's system goes beyond the scope of the data set, an in-house genetic algorithm package Fafoom[89] can be used to generate conformers and do the QM generation fast and automatically.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c01141.

Additional data that detail the parametrization process, force-field parameters, and analyses described in the paper (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Xiaojuan Hu − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*; Email: xhu@fhi-berlin.mpg.de

Kazi S. Amin − *Centre for Molecular Simulation and Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N 1N4, Canada*; Email: kazi.amin@ucalgary.ca

Dennis Salahub − *Centre for Molecular Simulation and Department of Chemistry, University of Calgary, Calgary, Alberta T2N 1N4, Canada*; Email: dsalahub@ucalgary.ca

Carsten Baldauf − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*; orcid.org/0000-0003-2637-6009; Email: baldauf@fhi-berlin.mpg.de

Authors

Markus Schneider − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*

Carmay Lim − *Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan; Department of Chemistry, National Tsing Hua University, Hsinchu 300, Taiwan*; orcid.org/0000-0001-9077-7769

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01141

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

The reference data can be found on the NOMAD repository via the DOI: 10.17172/NOMAD/2023.02.03-1.[72] FFAFFURR can be found at: https://github.com/XiaojuanHu/ffaffurr-dev/releases/tag/version1.0.[97] Code for implementing CTPOL in OpenMM: https://github.com/XiaojuanHu/CTPOL_MD.[71]

■ ACKNOWLEDGMENTS

The authors 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.

■ DEDICATION

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

■ 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 Proteomics* **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) George Cherian, M.; Jayasurya, A.; Bay, B.-H. Metallothionein in human tumors and potential carcinogenesis. *Mutat. Res.* **2004**, *533*, 201−209.

(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) 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.

(11) Wolfe, S. A.; Nekludova, L.; Pabo, C. O. DNA recognition by (Cys_2His_2) zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 183−212.

(12) 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.
(13) 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.
(14) Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.* **2012**, *137*, 230901.
(15) 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.
(16) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(17) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. 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.
(18) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 198−210.
(19) 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.
(20) 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.
(21) 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.
(22) Li, P.; Merz, K. M. Metal ion modeling using classical mechanics. *Chem. Rev.* **2017**, *117*, 1564−1686.
(23) 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.
(24) Maksimov, D.; Baldauf, C.; Rossi, M. The conformational space of a flexible amino acid at metallic surfaces. *Int. J. Quantum Chem.* **2021**, *121*, No. e26369.
(25) Schneider, M.; Baldauf, C. Relative energetics of acetyl-histidine protomers with and without Zn^{2+} and a benchmark of energy methods. 2018. arXiv preprint arXiv:1810.10596, https://arxiv.org/abs/1810.10596 (accessed on Oct 24, 2018).
(26) 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.
(27) 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.
(28) 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.
(29) Martinek, T.; Duboué-Dijon, E.; Timr, S.ˇ.; 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.
(30) 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.
(31) Li, P.; Song, L. F.; Merz, K. M., Jr Systematic parameterization of monovalent ions employing the nonbonded model. *J. Chem. Theory Comput.* **2015**, *11*, 1645−1657.
(32) Li, P.; Song, L. F.; Merz, K. M., Jr 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.
(33) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr 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.
(34) 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.
(35) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE-Antechamber python parser interface. *BMC Res. Notes* **2012**, *5*, 367−368.
(36) 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.
(37) 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.
(38) Kumar, A.; Yoluk, O.; MacKerell, A. D., Jr FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules. *J. Comput. Chem.* **2020**, *41*, 958−970.
(39) 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.
(40) 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.
(41) Tkatchenko, A.; DiStasio, R. A., Jr; Car, R.; Scheffler, M.; Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.
(42) 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*, 073005−073009.
(43) Gobre, V. V.; Tkatchenko, A. Scaling laws for van der Waals interactions in nanostructured materials. *Nat. Commun.* **2013**, *4*, 2341.
(44) 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.
(45) 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.
(46) Grimme, S. A general quantum mechanically derived force field (QMDFF) for molecules and condensed phase simulations. *J. Chem. Theory Comput.* **2014**, *10*, 4497−4514.
(47) Borodin, O. Polarizable force field development and molecular dynamics simulations of ionic liquids. *J. Phys. Chem. B* **2009**, *113*, 11463−11478.
(48) Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. Polarization effects in molecular mechanical force fields. *J. Phys.: Condens. Matter* **2009**, *21*, 333102.
(49) Allen, T. W.; Andersen, O. S.; Roux, B. Energetics of ion conduction through the gramicidin channel. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 117−122.
(50) 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.
(51) 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.

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

(52) 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.
(53) 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.
(54) 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.
(55) Soniat, M.; Rick, S. W. The effects of charge transfer on the aqueous solvation of ions. *J. Chem. Phys.* **2012**, *137*, 044511.
(56) 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.
(57) Friesner, R. A. Modeling polarization in proteins and protein−ligand complexes: Methods and preliminary results. *Adv. Protein Chem.* **2005**, *72*, 79−104.
(58) 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.
(59) Ponder, J. W.; Case, D. A. *Advances in Protein Chemistry*; Elsevier, 2003; Vol. *66*, pp 27−85.
(60) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.
(61) 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.
(62) Villa, F.; MacKerell, A. D., Jr; Roux, B.; Simonson, T.; 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.
(63) 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.
(64) 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.
(65) 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.
(66) Sakharov, D. V.; Lim, C. Zn protein simulations including charge transfer and local polarization effects. *J. Am. Chem. Soc.* **2005**, *127*, 4921−4929.
(67) 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.
(68) 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*, 160009−160013.
(69) 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*, No. 327.
(70) 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*, No. e1005659.
(71) Hu, X.; Baldauf, C. CTPOL_MD Source Code. https://github.com/XiaojuanHu/CTPOL_MD, (accessed on Dec 04, 2023).
(72) Hu, X.; Baldauf, C. FFAFFURR Data set. , (accessed on Dec 04, 2023).
(73) 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.
(74) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. Efficient integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **2009**, *228*, 8367−8379.
(75) 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.
(76) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(77) 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.
(78) 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*, 35772−35811.
(79) 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*, 144111.
(80) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129−138.
(81) Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.
(82) 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.
(83) 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.
(84) Wales, D. J.; Doye, J. P. K. 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.
(85) Wales, D. J.; Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **1999**, *285*, 1368−1372.
(86) 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.
(87) 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.
(88) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.
(89) Supady, A.; Blum, V.; Baldauf, C. First-principles molecular structure search with a genetic algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2338−2348.
(90) Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267−288.
(91) Hoerl, A. E.; Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55−67.
(92) Shi, Y.; Eberhart, R. C. Empirical study of particle swarm optimization *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*; IEEE, 1999, pp 1945−1950.
(93) 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.

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464

(94) 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.
(95) Lee, A. https://pythonhosted.org/pyswarm/(accessed on Dec 04, 2023).
(96) Dittner, M.; Müller, J.; Aktulga, H. M.; Hartke, B. Efficient global optimization of reactive force-field parameters. *J. Comput. Chem.* **2015**, *36*, 1550−1561.
(97) Hu, X.; Baldauf, C. *FFAFFURR* 1.0 Source Code. https://github.com/XiaojuanHu/ffaffurr-dev/releases/tag/version1.0 (accessed on Dec 04, 2023).
(98) 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.
(99) 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.
(100) 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.
(101) 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.
(102) Reilly, A. M.; Tkatchenko, A. van der Waals dispersion interactions in molecular materials: beyond pairwise additivity. *Chem. Sci.* **2015**, *6*, 3289−3301.
(103) 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.
(104) Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441−451.
(105) 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.
(106) 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.
(107) 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.
(108) 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.
(109) Li, P.; Merz, K. M., Jr 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. *Ann. Math. Stat.* **1962**, *33*, 1065−1076.
(111) Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **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 Res.* **2000**, *28*, 235−242.

Recommended by ACS

Developing an Implicit Solvation Machine Learning Model for Molecular Simulations of Ionic Media
Amaury Coste, Matej Praprotnik, *et al.*
DECEMBER 20, 2023
JOURNAL OF CHEMICAL THEORY AND COMPUTATION READ

Guidelines for Free-Energy Calculations Involving Charge Changes
Drazen Petrov, Vytautas Gapsys, *et al.*
JANUARY 02, 2024
JOURNAL OF CHEMICAL THEORY AND COMPUTATION READ

A Fast, Convenient, Polarizable Electrostatic Model for Molecular Dynamics
Liangyue Wang, Michael K. Gilson, *et al.*
JANUARY 19, 2024
JOURNAL OF CHEMICAL THEORY AND COMPUTATION READ

Variational Pair-Density Functional Theory: Dealing with Strong Correlation at the Protein Scale
Mikael Scott, Mickael G. Delcey, *et al.*
JANUARY 13, 2024
JOURNAL OF CHEMICAL THEORY AND COMPUTATION READ

Get More Suggestions >

https://doi.org/10.1021/acs.jctc.3c01141
J. Chem. Theory Comput. 2024, 20, 1448−1464