The Journal of Chemical Physics

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

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

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

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

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

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



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

## AFFILIATIONS

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

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

## ABSTRACT

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



## I. INTRODUCTION

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

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

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

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

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

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

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

## II. THEORY AND METHODS

### A. Cation–dipeptide reference structures

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

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

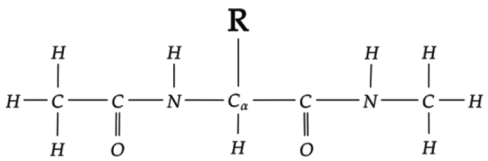


**FIG. 1**. Structure of a dipeptide, with a variable side chain (**R**) that extends from the alpha-carbon ($C_\alpha$), which can be any one of the 20 proteinogenic side chains, plus a few variations of His, namely, HSD (with hydrogen on $N^\delta$), HSE (with hydrogen on $N^\epsilon$), and HSP (with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

### B. Additive force fields for peptide–ion interactions

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

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

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

In Eq. (1), $K$, $b_0$, $\theta_0$, $S_0$, $\chi_0$, $n$, $\delta_n$, $\varepsilon_{ij}$, $\sigma_{ij}$, and $q$ are empirically determined parameters. The force constants [$K$ and parameters of the harmonic terms ($b_0$, $S_0$, $\theta_0$, $\chi_o$, $n$, and $\delta_n$)] are usually obtained using analysis of QM vibrational modes. The partial charges $q_i$ are generally obtained by fitting to electrostatic potential surfaces obtained via QM. After determining the bonded parameters and partial charges, the Lennard-Jones (LJ) terms ($\varepsilon_{ij}$, $\sigma_{ij}$) are finally fitted to reproduce both gas-phase QM energies and condensed-phase thermodynamics such as experimental hydration free energies.

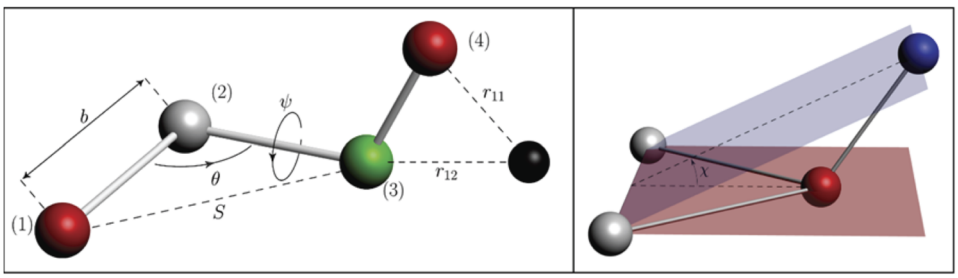


**FIG. 2**. Ball and stick model of classical FFs. Left: configuration of a proper dihedral. Right: configuration of an improper dihedral.

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

### C. The Drude polarizable force field

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

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

### D. Electrostatic interactions and polarization catastrophe

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

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

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

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

### E. Assessment of the cation–dipeptide interaction energies

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

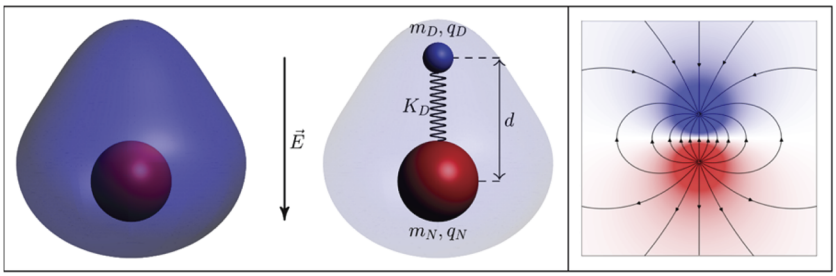


**FIG. 3**. Schematic of a Drude polarizable atom and resulting FF. Left: conceptual depiction of electron density around an atom polarized by an external electric field, $\vec{E}$. Middle: Drude model of the same atom, with Drude particles in blue. The Drude particles have a mass of $m_D$ and charge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

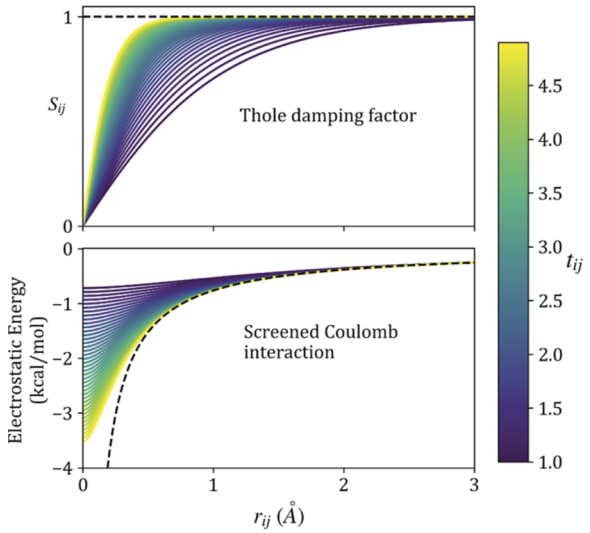


**FIG. 4**. Effect of Thole damping function for various $t_{ij}$. Each curve represents a particular screening factor ($t_{ij}$) where the color represents its value. The top panel shows $S_{ij}$ as a function of distance $r$, and the bottom panel shows the corresponding screened Coulomb potential energy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

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

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

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

## F. Charge transfer modeling with the CTPOL FF

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

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

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

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

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

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

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

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

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

## G. MD simulation protocol

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

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

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

## III. RESULTS AND DISCUSSION

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

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

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

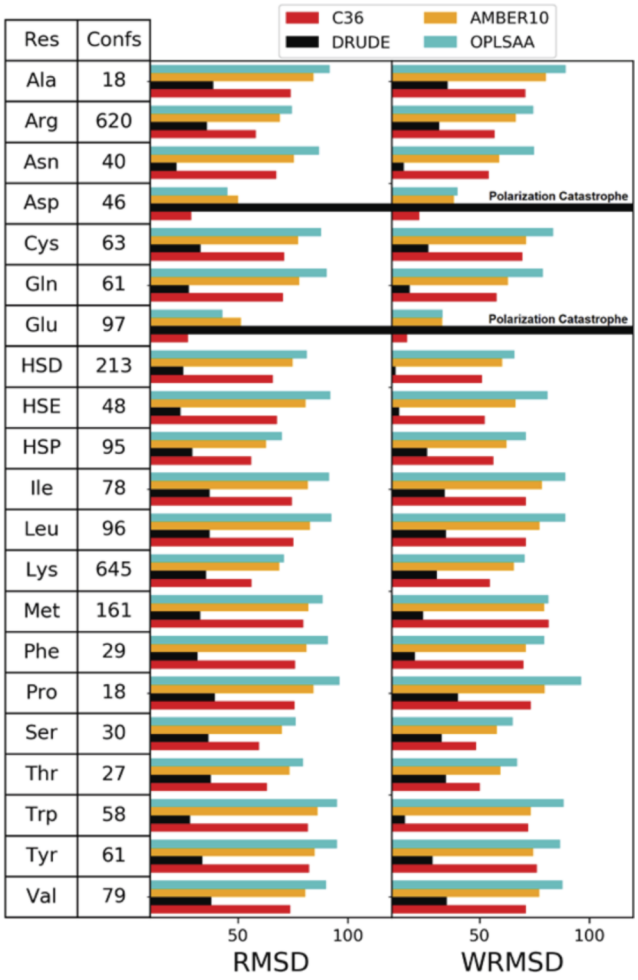


**FIG. 5**. Number of conformations and RMSD relative to the QM-interaction energies for each dipeptide residue. RMSDs are plotted for each of the four FFs prior to any optimization or correction. For Asp and Glu, the Drude RMSDs are on the order of $10^7$ due to polarization catastrophe. wRMSD is the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

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

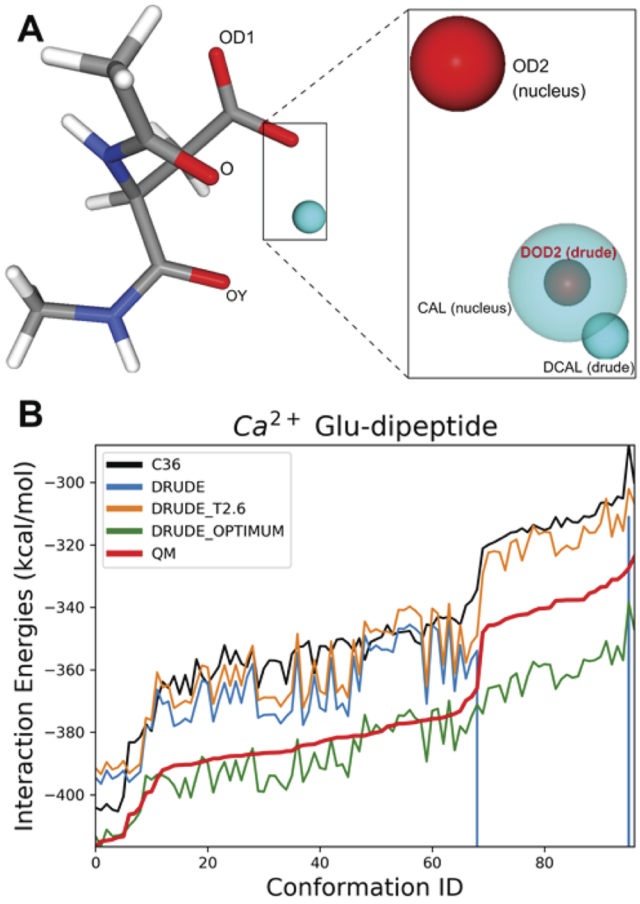


**FIG. 6**. (a) Polarization catastrophe in ASP dipeptide when $Ca^{2+}$ is in close proximity to the OD2 atom. The Drude particle of OD2 (atom-type DOD2) is abnormally pulled away from its parent nucleus by the electrostatic force of $Ca^{2+}$, which also has its Drude particle (DCAL) abnormally far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

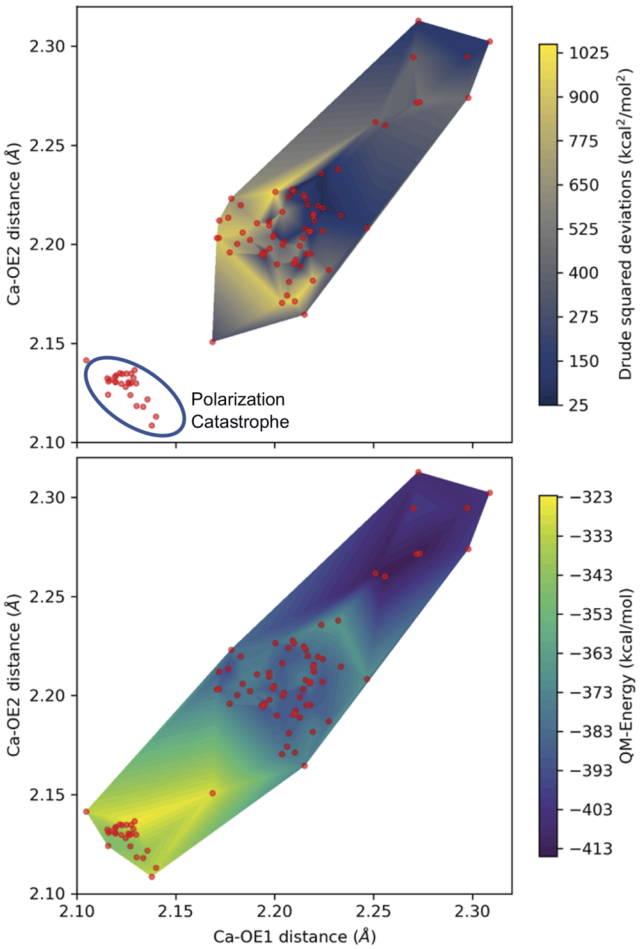


**FIG. 7**. (a) Drude FF squared energy deviations of Glu represented as functions of two collective variables—the distances in Å of $Ca^{2+}$ to OE1 and OE2 carboxylate oxygen atoms of Glu. Red dots represent each of the Glu conformations as a function of these two collective variables. Colors represent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

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

### B. Reduction of parameter space and avoiding polarization catastrophe

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

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

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

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

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

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

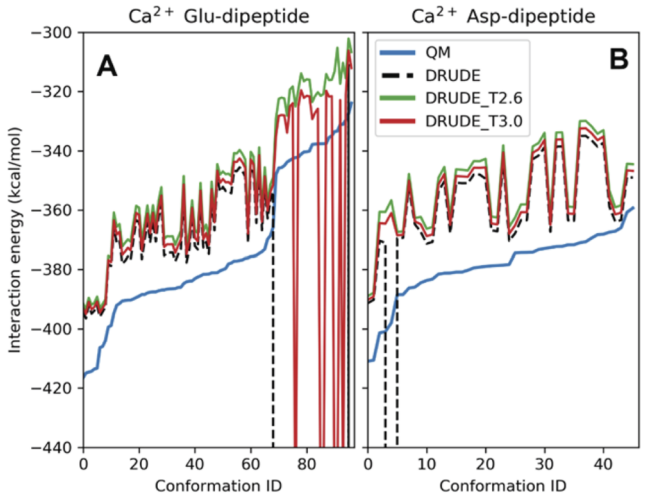


**FIG. 8**. Interaction energies of the Drude FF compared with QM for three values of the pair-specific Thole parameter for interaction between $Ca^{2+}$ (CALD) and Glu/Asp carboxylate oxygen (OD2C2A). The values of the Thole parameter are $t_{ij} = \infty$ (*dashed line*), 2.6 (*green*), and 3.0 (*red*),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

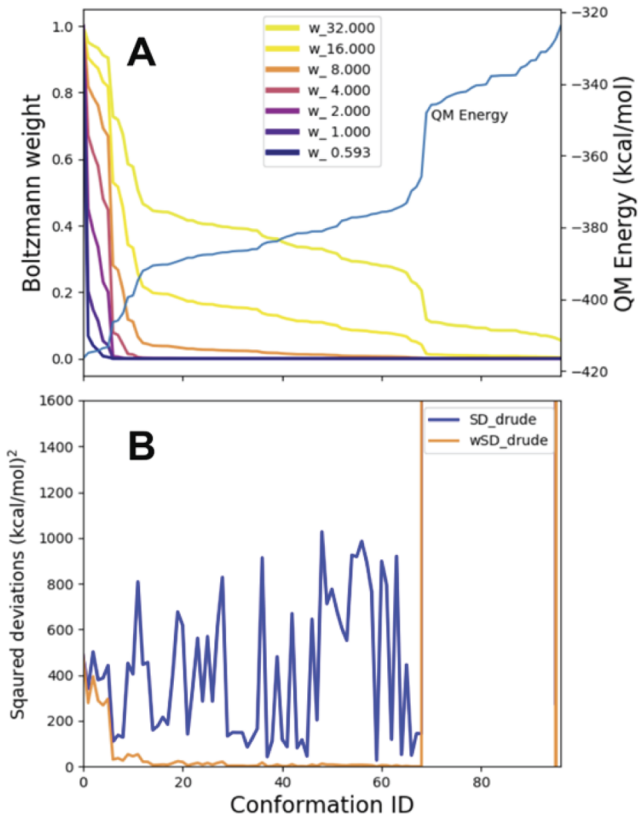


**FIG. 9**. Boltzmann weights applied to Glu-dipeptide–$Ca^{2+}$ interactions. (a) Boltzmann weight vs conformation ID at various RTs (0.593–32). The blue curve is the corresponding reference QM interaction energy. (b) Boltzmann-weighted squared deviations with RT = 8 (wSD) and unweighted squared deviations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

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

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

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

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

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

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

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

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

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

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

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

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

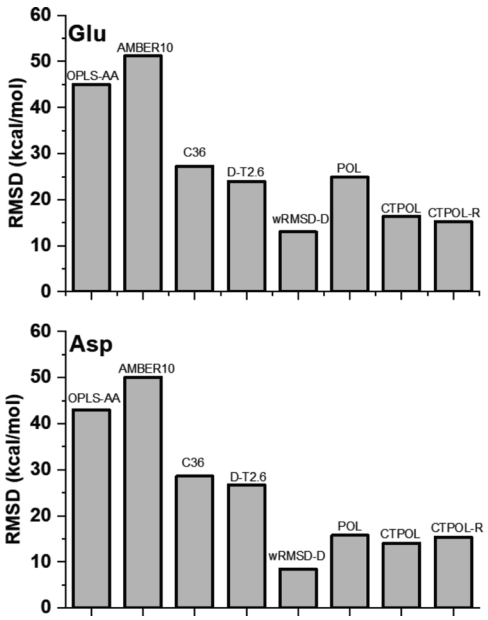


**FIG. 10**. RMSD estimated with Eq. (4) for the ensemble of conformations of Glu:$Ca^{2+}$ (top panel) and Asp:$Ca^{2+}$ (bottom panel) for various FFs used in our study. Drude data are shown for the Thole parameter set to 2.6 (D-T2.6) and optimized LJ and Thole parameters using Boltzmann-weighted RMSD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

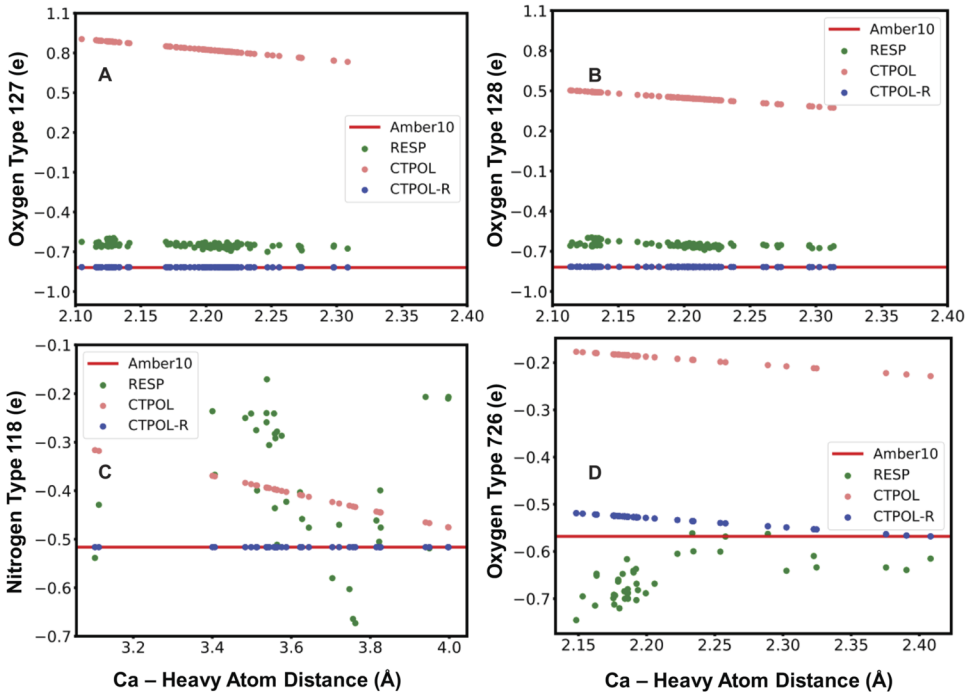


**FIG. 11**. Partial charge of ligand atoms in $Ca^{2+}$–Glu-dipeptide vs distance between an atom and $Ca^{2+}$. (a) and (b) show charges for the carboxylate oxygens (atom-types 127 and 128, as shown in Fig. SI 1). (c) and (d) show charge vs distance dependence for backbone nitrogen (atom-type 118)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

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

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

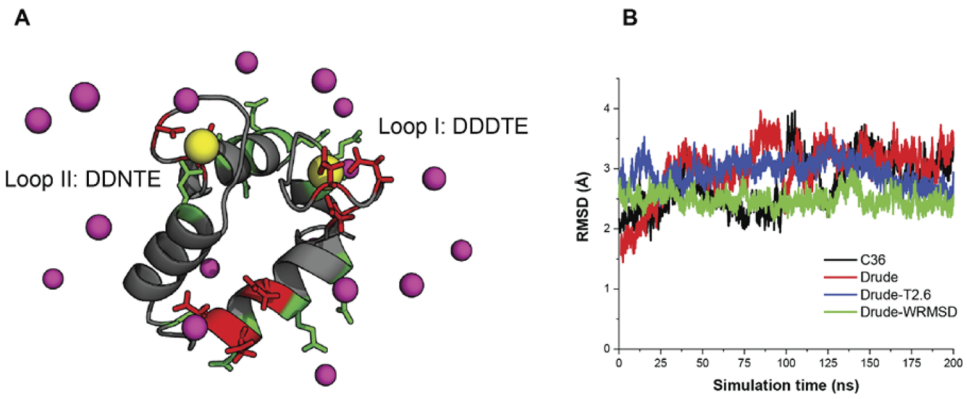


**FIG. 12**. (a) The N-lobe of the CaM protein with Loop I and II with two bound $Ca^{2+}$ ions (gold spheres). The positions of the aspartate residues are shown in red sticks, while glutamates are shown in green sticks. The $Ca^{2+}$ ions from the bulk solution are shown as magenta spheres. Water molecules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

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

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

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

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

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

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

### F. Evaluation of Drude-FF parameters in metalloprotein simulations

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

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

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

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

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

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

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

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

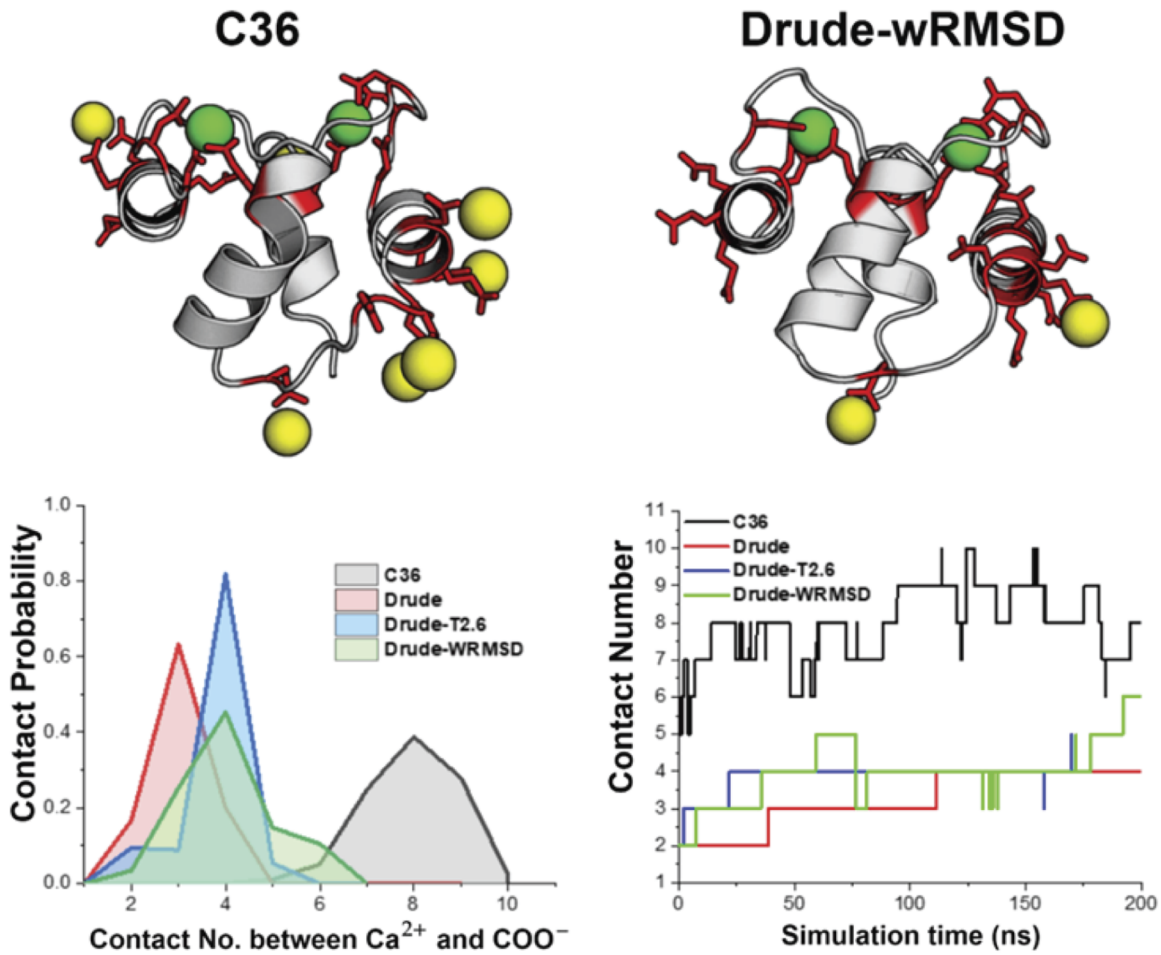


**FIG. 13**. (Top) Characteristic snapshots of a single EF-hand CaM in the solution with 150 mM $CaCl_2$ with contact numbers 9 and 4 observed in simulations with C36 (left) and Drude-wRMSD (right) FFs, respectively. In both snapshots, the two $Ca^{2+}$ ions bound to sites in Loop I and II are shown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

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

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

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

## IV. CONCLUSIONS AND OUTLOOK

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

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

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

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

## SUPPLEMENTARY MATERIAL

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

## ACKNOWLEDGMENTS

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

## AUTHORS' CONTRIBUTIONS

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

## DATA AVAILABILITY

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

## REFERENCES

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

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