Additive CHARMM36 Force Field for Nonstandard Amino Acids

Anastasia Croitoru, Sang-Jun Park, Anmol Kumar, Jumin Lee, Wonpil Im, Alexander Mackerell, Alexey Aleksandrov

Additive CHARMM36 Force Field for Nonstandard Amino Acids

Anastasia Croitoru[1], Sang-Jun Park[2], Anmol Kumar[3], Jumin Lee[2], Wonpil Im[2], Alexander D. MacKerell, Jr.[3*] and Alexey Aleksandrov[1*]

[1]Laboratoire d'Optique et Biosciences (CNRS UMR7645, INSERM U1182), Ecole Polytechnique, Institut polytechnique de Paris, F-91128 Palaiseau, France

[2]Departments of Biological Sciences, Chemistry, Bioengineering, and Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, USA

[3]Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, 20 Penn Street, Baltimore, Maryland 21201, USA

*Corresponding authors: alex@outerbanks.umaryland.edu and alexey.aleksandrov@polytechnique.edu

Running title: Additive CHARMM force field for nonstandard amino acids

Keywords: CHARMM General Force Field; CGenFF; force field development; post-translational modifications; artificial amino acids; chromophore;

ABSTRACT

Nonstandard amino acids are both abundant in nature, where they play a key role in various cellular processes, and can be synthesized in laboratories, for example, for the manufacture of a range of pharmaceutical agents. In this work we have extended the additive all-atom CHARMM36 and CHARMM General force field (CGenFF) to a large set of 333 nonstandard amino acids. These include both amino acids with nonstandard side chains, such as post translationally modified and artificial amino acids as well as amino acids with modified backbone groups, such as chromophores composed of several amino acids. Model compounds representative of the nonstandard amino acids were parametrized for protonation states that are likely at the physiological pH of 7 and, for some more common residues, in both D- and L-stereoisomers. Considering all protonation, tautomeric, and stereoisomeric forms, a total of 406 nonstandard amino acids were parametrized. Emphasis was placed on the quality of both intra- and intermolecular parameters. Partial charges were derived using quantum mechanical (QM) data on model compound dipole moments, electrostatic potentials, and interactions with water. Optimization of all intramolecular parameters, including torsion angle parameters, was performed against information from QM adiabatic potential energy surface (PES) scans. Special emphasis was put on the quality of terms corresponding to PES around rotatable dihedral angles. Validation of the force field was based on molecular dynamics simulations of 20 protein complexes containing different nonstandard amino acids. Overall, the presented parameters will allow for computational studies of a wide range of proteins containing nonstandard amino acids, including natural and artificial residues.

Introduction

Proteins are built from amino acids that are mostly incorporated biosynthetically into proteins during translation. The side chains of amino acids, defined by their distinct chemical characteristics, compose binding interfaces for partners in macromolecular complexes, create ligand binding sites, and assist chemical reactions occurring in enzyme catalytic sites. There are 20 amino acids in the standard genetic code and two additional amino acids that can be incorporated by special translation mechanisms.[1,2] Apart from these amino acids, however, there are many more nonstandard amino acids that are produced as a result of post-translational modifications (PTMs) in the cell, or can be synthesized and incorporated in laboratories.[3,4] PTMs of proteins significantly expand the chemical space, increase the complexity of the proteome, and play an important role in a wide range of functions in the cell.[5,6] PTMs not only can be incorporated by enzymes but also can arise as a consequence of oxidative stress.[7] Beyond natural ways of nonstandard amino acid incorporation, there has been a remarkable advance in the synthesis of nonstandard amino acids with novel characteristics and their incorporation into proteins.[3,8] Site-specific incorporation of nonstandard amino acids has been used to study protein structure, dynamics, and function by unique IR, X-ray, and fluorescent probes.[3,8] Furthermore, incorporation of nonstandard amino acids opened the door to novel biomaterials, enzymes,[9] and therapeutics.[10,11]

Molecular mechanics (MM) based simulation methods have become the most popular computational techniques for computational studies of biomolecular systems owing the system size and time-scale that can be accessed.[12,13] The major requirement for such computer simulations is the existence of a MM force field that defines the energies and forces acting on the molecular system. As such, the MM force field largely dictates the quality of these atomistic simulations. A number of force fields for nonstandard amino acids were derived previously and tools for the development of such force fields were reported.[14,15] AMBER parameters for 32 frequently occurring post-translational modifications were derived[16] which were later extended to include 147 noncanonical amino acids. Petrov et al. developed force field parameters for 256 different types of PTMs compatible with the GROMOS force field,[17] and later provided a web tool to incorporate PTMs into a 3D protein structure.[18] For the additive CHARMM force field a number of nonstandard amino acids were parametrized specifically in previous works.[19–21] 17 artificial amino acids were parametrized in our previous work.[15] CHARMM compatible topologies were created for 210 nonstandard alpha amino acid side chains[22] and were made available as an online service.[23] The set of the nonstandard amino acids included only amino acids that differ from the canonical amino acids by modifications in the side chains. These topologies and parameters for unknown functional groups were generated using the SwissParam web service,[24] which provides topologies based on the Merck molecular force field (MMFF).[25] However, no optimization of parameters was performed and the force field model is incompatible with the additive all-atom CHARMM36 force field.

The present study represents a systematic extension of the CHARMM36 additive force field to nonstandard amino acids,[26–29] also representing an extension of the additive CHARMM General Force Field (CGenFF) for small molecules.[30] The force field parameters, including charges and intramolecular parameters, were derived for the physiologically important protonation states and are of similar quality to those for the standard

amino acids. The parametrization method is based on the same protocol that is used to derive the CGenFF force field. The parametrization was done against quantum mechanical (QM) data, with a special emphasis on the dihedral terms corresponding to rotatable torsions. Results from MD simulations with the developed parameters of protein complexes containing nonstandard amino acids were then compared to the experimental structures for validation. To summarize, the extension of the CHARMM36 (C36) force field developed in this work is suitable to investigate interactions of nonstandard amino acids in the context of proteins.

METHODS AND MATERIALS

CHARMM potential energy function

The potential energy function of the non-polarizable all-atom CHARMM force field was adopted in this work for nonstandard amino acids.[12] This potential energy function is used for the remainder of the CHARMM36/CGenFF force field. The CHARMM potential energy is:

$U = U_{inter} + U_{intra}$ [Eq 1]

The intermolecular or non-bonded energy is due to electrostatic and van der Waals (vdW) interactions:

$$U_{inter} = \sum_{\substack{nonbonded \\ elec}} \frac{q_i q_j}{4\pi\varepsilon_0\, r_{ij}} + \sum_{\substack{nonbonded \\ vdW}} \varepsilon_{ij}\left(\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right)$$

[Eq 2]

The electrostatic term is described by Coulomb's law with q_i and q_j being the respective partial atomic charges on atoms *i* and *j*, and r_{ij} is the distance between atoms *i* and *j*. The vdW term is treated by the Lennard-Jones (LJ) 6-12 potential in which ε_{ij} is the well depth, $R_{min,ij}$ is the radius at which the LJ potential has a minimum. In the additive CHARMM force field, the LJ parameters for pairs of atoms *i* and *j* are constructed using the Lorentz-Berthelot combination rule.[31]

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \text{ and } R_{ij} = \frac{R_i + R_j}{2} \quad \text{[Eq 3]}$$

The intramolecular or bonded part of the potential energy function in Equation 1 is contributed by terms for the bonds, valence angles, dihedral angles, improper dihedral angles, and selected Urey−Bradley terms. In addition, the bonded energy function has been extended to include the CMAP cross-term applied to improve the conformational properties associated with the ϕ and ψ torsion angles of the peptide backbone. The intramolecular part is given by:

$$U_{intra} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{Urey-Bradley} K_{UB}(r_{1-3} - r_{1-3;0})^2 + \sum_{dihedral} \sum_{n=1}^{N} K_n\,(1 + \cos(n\varphi - \delta_n)) + \sum_{improper} K_\varphi(\varphi - \varphi_0)^2 + CMAP$$

, [Eq 4]

where b_0, θ_0, $r_{1-3;0}$, and φ_0 are the bond, angle, Urey−Bradley, and improper dihedral angle equilibrium values, respectively; the K's are the force constants; and n and δ_n are the dihedral multiplicity and phase. A dihedral term is represented as a Fourier series with N number of multiplicities, and the CMAP term is a special grid-based dihedral correction map applied to the protein backbone.[27] The current CHARMM force field uses less than seven multiplicities (N<7) for a dihedral term with only two possible values for phases: 0º or 180º. An improper dihedral angle is defined between four atoms, but in contrast to the dihedral angle, three of the atoms are bonded to the central atom and in the CHARMM force field φ_0 is typically set to zero.

Parametrization Protocol

The atom types were adapted from CGenFF.[30] The ParamChem web server (https://cgenff.umaryland.edu/) was used to assign existing atomic types and to obtain initial guesses of the partial atomic charges and bonded parameters for the model compounds.[32,33] Partial charges that were assigned a zero penalty by ParamChem, i.e., already optimized in CGenFF, were not considered for optimization in the present study with the exception of selected zero-penalty atoms covalently-linked to high-penalty atoms. Parameters of the LJ potential were taken from the CGenFF force field and were not further optimized in this work. We use the atom names from the Protein Data Bank for non-hydrogen atoms in residues, which itself uses the convention defined in the PDB Chemical Component Dictionary (CCD).[34,35] The atom names for hydrogens were assigned according to the parent heavy atom to which they are bonded. The parametrization protocol is shown in Figure 1.

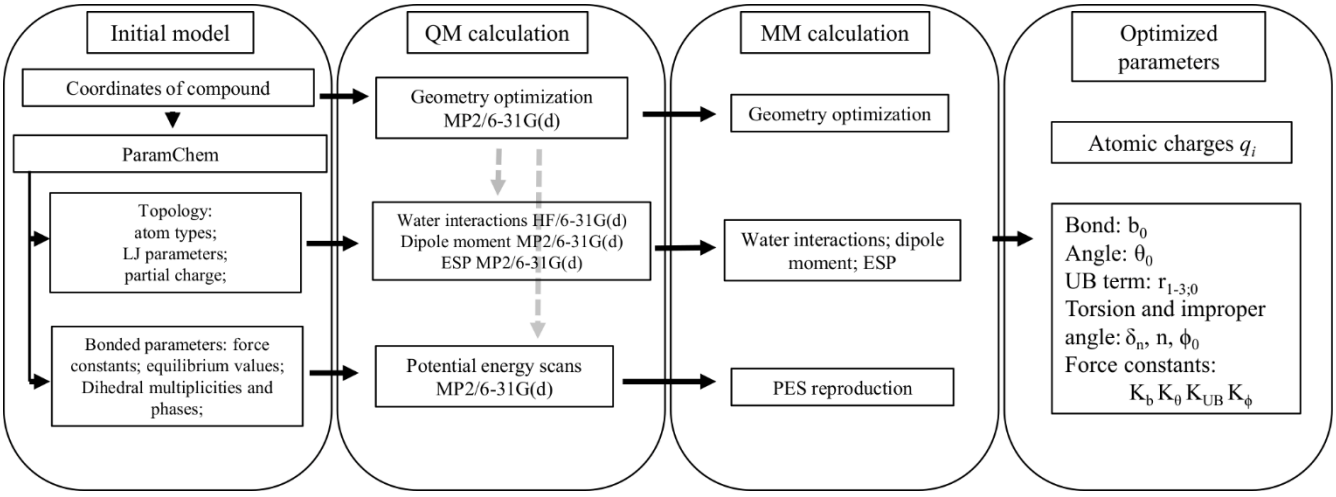

Figure 1. Workflow of force field parametrization. For anionic species, the MP2/6-311G(d) model chemistry was used for optimization and potential energy surface (PES) scans of the compounds.

Given a model compound, the protocol starts by defining bonded parameters that need to be optimized for this molecule. These parameters are those that do not explicitly exist in CGenFF, but are assigned based on analogy with known CGenFF parameters by the CGenFF program (see below). The initial geometry for the model compound is constructed using the available PDB coordinates for the corresponding nonstandard amino acid and

by adding protons using the Babel software.[36] The geometry is further optimized at the MP2/6-31G(d) model chemistry, or MP2/6-311G(d) model chemistry for anionic molecules. The resulting QM geometry is then used to optimize atomic charges as described below. The MM model with optimized charges is used to optimize bonded terms in the next step. Adiabatic potential energy scans with QM, discussed in detail below, are performed along the degrees of freedom for high-penalty parameters, which are those not explicitly present in CGenFF. Those bonded parameters are optimized to minimize differences between QM and MM geometries and potential energy surfaces. In this work, we first optimized terms associated with dihedral angles including soft dihedral angles, along which large conformational fluctuations are possible; then bonded terms associated with other degrees of freedom were adjusted. The steps were repeated iteratively at least two times, and the optimization was stopped when no significant improvement was obtained in further iterations.

Choice of atom types and model compounds

The nonstandard amino acids parametrized in this work represent a broad and heterogeneous set of molecules. The set of nonstandard amino acids was divided into two groups of residues depending on the need to parametrize the backbone group. For the residues of the first group the backbone atom types and associated parameters from the C36 force field are used, while the side chains up to the Cβ atom have CGenFF atom types. The terms corresponding to the bond Cα-Cβ between the backbone group and side chain have both CGenFF and C36 atom types. This allows the use of well-developed parameters from the C36 force field for the backbone of these residues including the CMAP term. The amino acids with backbone groups different from the backbone of the standard amino acids have CGenFF atom types and parameters for all atoms of the nonstandard amino acid. For the residues in this group the bonded terms corresponding to the peptide bonds between the nonstandard residues and neighboring residues have both CGenFF and C36 atom types as represented in Figure S1. The CMAP term was not included for this class of residues; however, all dihedral angles including those associated with the backbone atoms were carefully parametrized using potential energy surface (PES) scans.

Charges and bonded parameters for nonstandard amino acids were optimized using model compounds. In the charge optimization for amino acids with the standard backbone atom types, the model compound included the side chain group up to Cα or Cβ to parametrize the bonded terms associated with the side chain. However, if it was possible, smaller compounds were used, and several compounds were included for large side chains. Such amino acids were further broken down into several parts with the cleavage sites chosen between two acyclic saturated carbons. A proton was added to the acyclic saturated carbon of the cleavage site to complete the chemical structure of the model compound. All the nonstandard amino acids and the associated model compounds are presented in the Supporting Information. For the bonded terms of the amino acids with the standard backbone group, the torsion terms corresponding to the rotation around the bond Cα-Cβ (χ1) were optimized using dipeptides as model compounds, which represented a modified residue with acetylated N-terminus and N-methylamide C-terminus. Dihedral angles φ and ψ of the backbone were constrained to −60° and −45°, respectively, corresponding to the ideal values in an α-helix. For nonstandard amino acids with backbone groups different from the standard

backbone, tetrapeptides were used to optimize the parameters corresponding to the peptide bonds between the nonstandard residues and neighboring residues. The tetrapeptides had the sequence ALA-X-ALA with acetylated N-terminus and N-methylamide C-terminus, where X is a nonstandard amino acid, with the backbone groups of the flanking residues constrained to the ideal α-helix geometry.

Determination of the intermolecular force field parameters

The intermolecular energy is due to Coulomb and LJ terms. Consistent with the development of the CHARMM force field, atomic charges were optimized targeting interactions between the model compound and individual water molecules, and the dipole moment of the model compound. Quantum mechanical electrostatic potentials (ESP) have also been used as additional target data in the charge fitting similar to the other work.[37] However, the weighting of the ESPs was smaller than that used for water interactions (see below). The charge optimization was performed on the compound structures optimized with the MP2 level of theory[38] and 6-31G(d) basis set[39] and 6-311G(d) for anionic molecules. Gaussian09[40] was used for all QM calculations. All QM optimizations were performed to default tight tolerances. Since optimization is performed in vacuum, for model compounds containing carboxylic acid and amine fragments, and that can exist in zwitterionic forms in aqueous solvent, the distance between protons on the amine group and the amine nitrogen was constrained to prevent protonation of the carboxylic group by proton transfer from the protonated amine group.

Atoms of the model compound that can participate in hydrogen bonds were probed by individual water molecules placed in idealized linear orientations.[26] Different orientations of the water molecule were considered around the interaction axis: the complex was calculated every 45º or 90º of the water probe rotation for polar atoms, and one or two orientations for non-polar atoms. All model compound-water interaction orientations are presented in the Supporting Information. Each water-model compound complex was optimized by varying the interaction distance between the water and the model compound with the monomer geometries fixed to find the minimum interaction energy distance. The QM-optimized gas-phase geometry was used for the model compounds as described above, and TIP3P model geometry for the water molecule. The angle defining the orientation of the water molecule around the interacting axis was held fixed during the optimization. The interaction energy was calculated for the minimum interaction energy distance. Calculations were done at the HF/6-31G(d) level.[26,30] Following the CHARMM standard protocol, the *ab initio* interaction energies were scaled (made more favorable) by an empirical factor of 1.16 only for neutral polar compounds and the HF/6-31G(d) minimum interaction distance was corrected by subtracting 0.2 Å for all polar interactions involving neutral compounds.[26] In the case of sulfur atoms, the model compound-water interactions were calculated at the MP2/6-31G(d) level including the basis set superposition error (BSSE) correction of Boys and Bernardi[41] and without applying standard scaling and offset rules.

The molecular dipole moment, which is defined by the charge distribution, was used to provide additional target data for the optimization of the atomic charges. The dipole moment was included only for the neutral compounds in the charge fitting.[42] The dipole moment was calculated in vacuum at the MP2/6-31G(d) model

chemistry using the QM-optimized conformation.[30] Following the standard CHARMM protocol, to account for the molecular polarizability implicitly, the MM optimization targeted dipole moments increased by 30% with respect to the QM values.[30] Both the magnitude and direction of the dipole moment were targeted.[42]

QM water interaction data may not be sufficient to define partial charges on all atoms for large compounds, since only water interactions with a few hydrogen-bond donors and acceptors at the molecular surface are probed. Therefore, ESP calculations were performed at the MP2/6-31G(d) model chemistry and at MP2/6-311G(d) for anions,[37] with the resulting ESPs used in the charge optimization to facilitate the determination of charges on atoms not involved in hydrogen bond interactions with water. At each iteration during the charge optimization, the root mean square deviation (RMSD) between QM and MM ESPs was evaluated and added with the corresponding weight to the target function. However, the weight for the ESPs (the corresponding weight: 1.0 $kcal^{-1}\cdot mol\cdot Å$) was kept small relative to the weights for water-interaction (10.0 $kcal^{-1}\cdot mol$) and dipole moment contributions (3.0 D^{-1}), as the reproduction of water-compound interaction energies and geometries is important to balance the solvent-solvent, solvent-solute, and solute-solute interactions.

The charge optimization was performed with the C++ program that was used to parametrize a large number of modified nucleotides in our previous work.[42] The following terms were included with different weights in the target function: the RMS deviation between empirical and *ab initio* minimum interaction energies, the RMS deviation between *ab initio* and empirical minimum interaction distances, the absolute difference between the norms of the empirical and *ab initio* dipole moments, the angle between the empirical and *ab initio* dipole moments, the RMS deviation between *ab initio* and empirical ESPs, and a term associated with restraints on the charges. The latter term was introduced to prevent large deviations from the starting guess for the charges. Charges of symmetrical atoms had identical values during the charge optimization. The initial partial charges were obtained from the ParamChem online server (https://cgenff.umaryland.edu/). Charges that were already optimized in CGenFF, for example for benzene, were not further adjusted in this work. Charges of aliphatic hydrogen atoms were not optimized, in accord with the standard CHARMM method with aliphatic hydrogen atoms having a charge of +0.09*e*. The LJ parameters were not considered for optimization. For seven complex model compounds two local minimum geometries were used simultaneously in charge fitting. In each geometry different hydrogen-bond sites were probed by water interactions, which are not accessible in the other geometry due to interactions with other groups of the compound.

Optimization of flexible dihedral parameters

Dihedrals within a molecule can be classed in two groups, soft or rotatable versus stiff or non-rotatable. PES associated with non-rotatable dihedrals (e.g., dihedral angles about double bonds or in ring systems) are typically characterized by a single minimum and high energy for small deformations. Rotatable dihedrals have a shallow energy surface with relative small barriers between minima and, thus, may undergo large fluctuations during simulations. Since the molecule can undergo large conformational motions along rotatable dihedrals, accurate treatment of these dihedral terms is paramount. Each compound has 1 to N, $\{\chi_i\}$, rotatable dihedrals. To

parametrize these terms adiabatic PES scans were performed for each torsion, χ_k , in which the torsion angle was scanned in the range from -180° to 180° in 10° increments. During these scan calculations, the compound was energetically optimized along all degrees of freedom, except for the soft dihedral angles. The scanned soft dihedral χ_k was constrained to the target value, while all other soft dihedrals $\{\chi_{i\neq k}\}$ were constrained to the values corresponding to the minimum energy geometry of the model compound. QM calculations were performed at the MP2/6-31G(d) model chemistry (MP2/6-311G(d) for anions). Each conformation for the MM calculations was extracted from the QM scan and minimized with a harmonic restraint with the force constant of $5 \cdot 10^4$ kcal·mol^{-1}·radian^{-2} on the target torsion. All other rotatable dihedrals were restrained with the same force constant to the values corresponding to the minimum energy geometry. Using these dihedral restraints, we ensure that the QM and MM structures for each dihedral PES scan are close to each other, i.e. that we compare the same region on QM and MM PES surfaces. The dihedral parameters were optimized to achieve a minimum deviation between the QM and MM surfaces only in the low-energy regions with energies <10 kcal·mol^{-1} above the minimum energy.

Optimization of bonded harmonic energy terms

Parameters for the intramolecular terms described by harmonic potentials; bonds, valence angles, Urey-Bradley terms, and improper dihedrals, as well as non-rotatable dihedral angles were optimized using the following protocol. The initial guess for force constants was provided by the ParamChem online server as described above. The initial equilibrium values for bonds, valence angles, and Urey-Bradley distances were taken directly from MP2/6-31G(d) geometries (MP2/6-311G(d) for anions). Only parameters with the ParamChem penalty >10 were considered for optimization. The equilibrium angle for improper terms was set to zero and was not optimized. An adiabatic PES scan for each degree of freedom that has adjustable parameters in the force field was performed. The same method was also used in CGenFF to determine force constants by three-point PES scans, when the assignment of contributions of the internal coordinates to the vibrations was ambiguous.[42,43] During the PES scans performed by varying one stiff degree of freedom the potential energy may become very high, even for relatively small deformations. Such high-energy regions of PES are not sampled during typical MD simulations. To ensure that only relevant regions of PES are parametrized we use the method from our previous work[44] to limit deformations and corresponding energies. In this method using initial values for distortions, force constants of energy terms are estimated. The initial values for the distortions are then corrected using the following equation:

$$\Delta x' = \sqrt{2\,\Delta E_{max}/k}, \qquad \text{[Eq 5]}$$

where $k = 2\,(E(\Delta x) - E_0)/\Delta x^{\,2}$. Δx and $\Delta x'$ are the initial and adjusted maximum distortions, respectively; E_0 and $E(\Delta x)$ are the minimum energy and energy of the deformed structure. ΔE_{max} defines the highest energy on scanned PES. To optimize each bonded term seven points were used on PES equally spaced in the range of $x \in [x_0 - \Delta x',\, x_0 + \Delta x']$, including the minimum energy structure at $x = x_0$. 2.0 kcal·mol^{-1} was used for ΔE_{max} in Equation 5. All PES scans were performed at the MP2/6-31G(d) model chemistry and MP2/6-311G(d) for anions.

The equilibrium values of the MM parameters and force constants were adjusted simultaneously using a C++ program based on the Powell minimization algorithms from Numerical Recipes.[45] Each conformation for the

MM calculation was extracted from the QM scan and minimized with a harmonic restraint force constant of $5 \cdot 10^4$ kcal·mol^{-1}·Å^{-2} or $5 \cdot 10^4$ kcal·mol^{-1}·radian^{-2} on the target bond and valence angle, respectively. At each optimization iteration of bonded parameters, PES adiabatic scans were performed with CHARMM using a new set of MM parameters. The target function included RMS deviation between QM and empirical PES energies; QM and MM geometries; and restraints to the initial set of parameters provided by the ParamChem server. In addition, the weighted RMS deviation between Cartesian components of QM and CHARMM forces was added to the target function. The MM parameters were adjusted until the target function could not be reduced further. MM calculations were performed with the CHARMM program.[46]

Molecular Dynamics simulations

To evaluate the quality of the force field model for nonstandard amino acids, molecular dynamics (MD) simulations of 20 protein complexes were performed. The protein complexes are summarized in Table S1. The crystal structures with a high to medium resolution were retrieved from the Protein Data Bank (PDB). Each system contained all protein residues for small and medium size proteins and a spherical truncated model centered on the modified residue was used for large protein complexes. Protonation states of residues were assigned using PROPKA,[47,48] while protonation states of histidines were assigned by visual inspection and ideal stereochemistry. In addition to crystal waters, a cubic box of water was overlaid and waters overlapping the protein and crystal water molecules were removed based on a minimum distance of 3.5 Å between non-hydrogen atoms. The size of the water box was chosen so that the shortest distance between protein atoms and the box edges was 10 Å. Periodic boundary conditions were assumed and all long range electrostatic interactions were computed efficiently by the particle mesh Ewald method[49] using a real-space cutoff of 11 Å. The appropriate number of potassium or chloride counterions was included to render the system electrically neutral. A smooth switching function was used to truncate all van der Waals interactions at the distance of 11 Å. Long range electrostatic forces were evaluated every 4 steps, while short-range non-bonded interactions were computed at each step. MD simulations were performed at constant room temperature and pressure, after 200 ps of thermalization. Constant pressure was maintained using the Berendsen pressure bath coupling[50] with the relaxation of 500 fs, the compressibility parameter of liquid water. Constant temperature was maintained by coupling to a heat bath with a room temperature by correcting forces as implemented in the NAMD program.[51] For truncated protein systems, the simulation setup was similar to previous studies.[52,53] In brief, the simulations included protein residues within a 24 Å sphere around the nonstandard amino acid. Protein atoms between 20 and 24 Å from the sphere's center were harmonically restrained to their experimentally determined positions. The CHARMM36m force field was used for the protein[28,37] and the TIP3P model for water.[26,54,55] The nonstandard amino acid was modeled using the force field parameters specifically developed in this work. Calculations were done with the NAMD program running on GPUs for efficiency.[51] MD simulations of the protein complexes were continued for 100 nanoseconds.

Results and Discussion

Set of parametrized molecules

In this work a total of 333 nonstandard amino acids were parametrized. Chemical structures and amino acid names are given in Figure S2 and Table S2 in Supporting Information. This set of residues includes 198 amino acids from the SwissSidechain database of nonstandard amino acids.[23] In addition, another 134 frequent nonstandard amino acids were considered, including 42 nonstandard amino acids with modified backbone moieties. The D- and L-stereoisomers were considered for 61 residues. To designate D-stereoisomers, the letter D was added at the beginning of three letter code of the residue. The pK_a's and tautomeric states were predicted with MarvinSketch software version 19.19.[56] The most important protonation and tautomeric states at the physiological pH of 7 were considered. We use the three letter code for deprotonated forms of residues and the four letter code with the letter P at the end to designate the protonated form. These residues are: TPQ (TPQP), PHD (PHDP), MHS (MHSP), LLP (LLPP), IT1 (IT1P), HIC (HICP), DDE (DDEP), CYQ (CYQP), CGU (CGUP), and GGB (GGBP). For 2-fluoro-L-histidine (residue name: 2HF), two tautomeric forms were considered for the neutral state: protonated on Nε and protonated on Nδ, named 2HFE and 2HFD, respectively. Four amino acids, AYA, CXM, FME, and PR4 are present at the N-terminus as they appear in the PDB structures only in N-termini. Two residues, C2N and FLA, are present in the force field model only as standalone ligands as they are not present in polypeptides in the PDB. The set of amino acids with the standard C36 backbone group includes 358 residues and the set of amino acids with nonstandard backbone groups includes 42 residues. Overall, considering all protonation, tautomeric, and stereoisomeric forms, 406 nonstandard amino acids were parametrized based on a total of 188 model compounds.

Charge optimization

The CHARMM partial charges were derived targeting water-compound interactions, the dipole moment magnitude and its orientation, and ESP. The amino acids were broken down into smaller compounds as described in the Methods section, giving 188 model compounds that were not previously optimized in the CGenFF force field and required charge optimization. The model compounds include 52 ionized compounds and 136 neutral compounds. Atomic charges of these molecules were further optimized. One QM minimum-energy geometry was considered for 181 model compounds and two local-minimum geometries were considered for seven complex molecules. A total of 3857 monohydrate probe water-model compound interaction complexes were used as target data as explained in Methods. This includes 906 probe water-model compound interactions for ionized and 2951 for neutral compounds.

Table 1. Statistics for intermolecular parameter development and agreement with respect to selected target data for all model compounds used to parametrize nonstandard amino acids.

Property	*N* points	RMSD	MAE
		optimal/initial	optimal/initial
[a]norm of μ	141	0.59/1.72	0.41/1.28

[b]direction of μ	141	2.4/16.3	5.1/32.7
[c]water-solute E_{int}	3857	0.46/1.79	0.32/0.95
[d]water-solute d_{min}	3857	0.20/0.66	0.16/0.28
[e]φ_{elec}	195	2.38/4.19	1.96/3.39

[a]The magnitude of the dipole moment (μ) is given in D; [b]angle (°) between the *ab initio* and empirical dipole moment vectors, the numbers in the RMSD and MAE columns correspond to the average angle and the average dipole moment-weighted angle (using $\sum \varphi_i \cdot p_i / \sum p_i$ where p_i is the magnitude of the QM dipole moment and φ_i is the angle between the MM and QM dipole moments), respectively; [c,d]probe water-model compound interaction energies and distances are in kcal·mol^{-1} and Å, respectively; [e]electrostatic potential is in kcal·mol^{-1}·Å^{-1}.

Figure 2 compares QM and MM interactions energies. The statistics for water-compound interactions for all compounds is given in Table 1. Empirical and *ab initio* interaction energies and distances are given in Table S3-S678 in Supplementary Information. The RMS deviation for interaction energies with the initial ParamChem and optimized charges is 1.79 kcal·mol^{-1} and 0.46 kcal·mol^{-1}, respectively, while the mean absolute error (MAE) is 0.95 kcal·mol^{-1} and 0.32 kcal·mol^{-1}, respectively. In this work, several probe water orientations were considered for a compound atom that can participate in H-bonds, in contrast to the C36 force field where usually one interaction was considered to probe each atomic site in the molecule. Some of these orientations have much higher interaction energies due to interactions with other groups in the molecule, and are more difficult to reproduce by the simple additive form of the force field. This explains why the RMS deviation between QM and MM interaction energies of 0.46 kcal·mol^{-1} obtained in this work is slightly higher relative to 0.34 kcal·mol^{-1} reported for the CGenFF force field.[30] The initial ParamChem charges assigned by analogy systematically overestimate interaction energies with probe water molecules in Figure 2 by 6%. However, with ParamChem charges interactions can be significantly underestimated or overestimated as demonstrated in Figure 2, if the analogous groups are not available in CGenFF. The slope for the interaction energies computed with both the initial guess and optimal charges is close to one, demonstrating that the force field model can reproduce solvent interactions for a wide range of the nonstandard amino acids. The RMS deviation for minimum-energy interaction distances is 0.28 Å with the initial ParamChem guess, which decreased to 0.16 Å with the optimized atomic charges. The agreement for interaction distances is comparable to that previously reported for CGenFF with the distance RMS deviation

of 0.20 Å.[30]

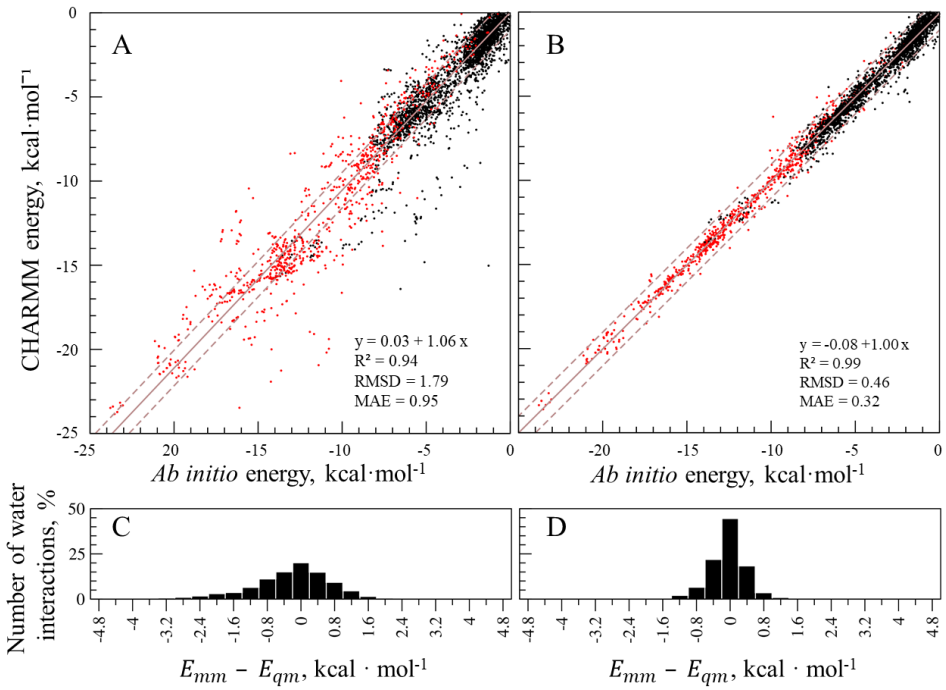

Figure 2. Corrected QM and CHARMM water interaction energies for the compound-water monohydrates. A) The CHARMM energies were computed using the initial ParamChem charges and B) the optimized atomic charges. C) and D) Percentage of water interactions vs energy deviation in A) and B), respectively. Interaction energies are shown in red and black for ionized and neutral compounds, respectively. The linear regression line between QM and CHARMM data is shown by the solid line. The diagonal dashed lines represent deviations of ±1.0 kcal·mol^{-1} from the regression line.

The statistics for empirical and *ab initio* dipole moments are given in Table 1. The dipole moment was included only for neutral compounds consistent with the standard CHARMM protocol. The CHARMM additive force field charge optimization targets systematically overestimated interactions with water to implicitly include the contribution of electronic polarization of molecules in an aqueous environment. Consistent with this, the empirical dipole moments should overestimate the gas phase dipole moments by ~30%.[30] The initial ParamChem charges yield dipole moments that are within 1 D of the target scaled QM values for the majority of compounds, though significant deviations are present in a number of cases (Figure 3 and Tables S3-S678). The dipole moments with the optimal charges are significantly improved relative to the dipole moments computed using the initial set of ParamChem charges. The RMS deviation between scaled QM and MM dipole moments averaged over all model compounds is 1.7 and 0.6 D computed with the initial ParamChem and optimal charges, respectively. The orientation of the dipole moment is also improved, and the angle between the QM and MM dipole moment averaged over the neutral model compounds is 32.7º and 5.0º with the initial and optimal set of charges, respectively. The RMSD for the dipole moment direction in this work is comparable to or better than the agreement of 8.5º obtained for the original CGenFF force field.[30] The angle between the QM and MM dipole moments for all model compounds except for three cases is smaller than 10º, and larger than 10º only for three molecules with a

very small dipole moment (<0.5 D). Consistent with this, the average dipole moment-weighted angle between the QM and MM dipole moments (computed using $\sum \varphi_i \cdot p_i / \sum p_i$ where p_i is the magnitude of the QM dipole moment and φ_i is the angle between the MM and QM dipole moments) is 16.3º and 2.4º with the initial and optimal charges.

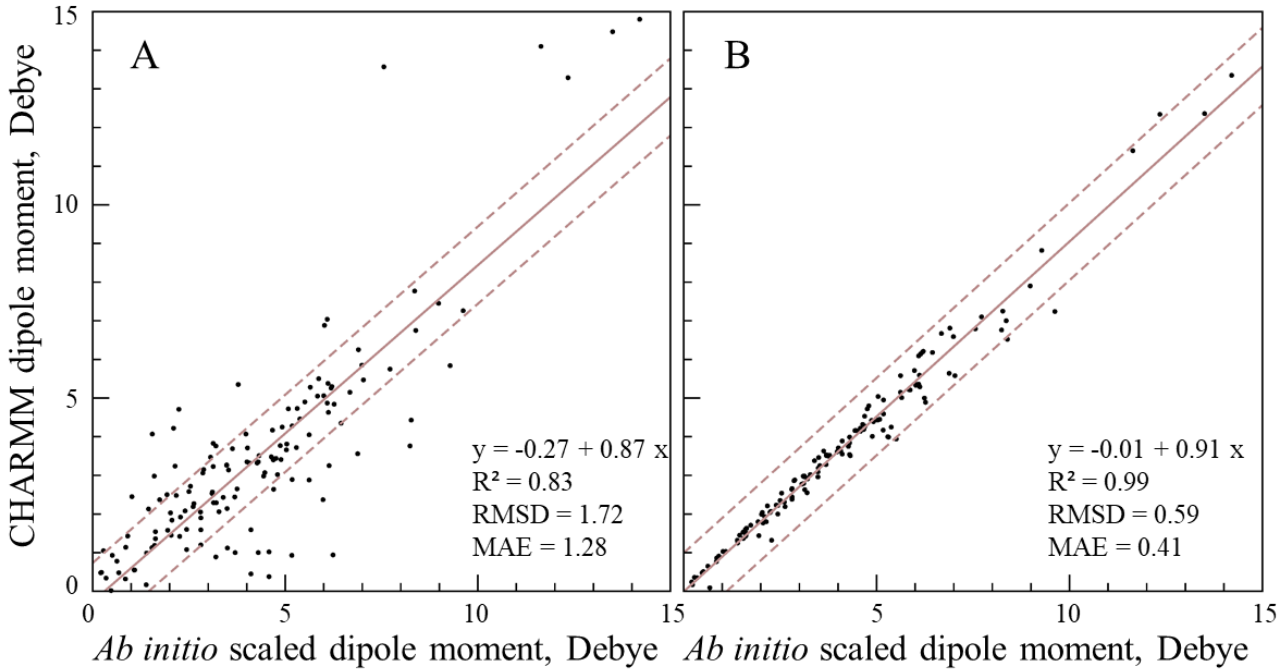

Figure 3. Comparison between the scaled QM (increased by 30%) and CHARMM dipole moments for 141 neutral model compounds. The CHARMM dipole moment was computed using A) the initial ParamChem charges and B) using the optimized atomic charges. The linear regression line is shown by the solid line; the dashed lines represent deviations of ±1 D from the regression line.

ESPs were included as an additional restraint to provide better charge distribution in the model compound as in a previous study.[37] However, the weight for the ESP potential was weak to achieve a better agreement for the water interactions. Nonetheless, for all model compounds, including ionized molecules, the ESPs are significantly improved relative to the initial values. The relative number of molecules vs ESP RMS deviation with the initial and optimal set of charges is shown in Figure S3. The RMS deviation between MM and QM electrostatic potential averaged over 195 compounds and geometries is 4.2 and 2.4 kcal·mol^{-1}·Å^{-1} with the initial ParamChem and optimal set of charges, respectively. Targeting the QM ESP was found to be particularly important for ionized compounds, since the number of probe water interactions was fewer than for neutral compounds, due to the dominant contribution of the net charge to water interactions, as well as to the lack of the inclusion of dipole moments as target data.

The largest absolute difference between the initial ParamChem and optimized charges was observed for atoms in residues SUN (O-[(R)-(dimethylamino)(ethoxy)phosphoryl]-L-serine) and SVX (O-[(R)-

ethoxy(methyl)phosphoryl]-L-serine). Both residues are similar: instead of the methyl group in SVX, SUN has the dimethylamino group bonded the phosphorous atom. In both cases, the largest difference ($q_{initial}$-$q_{optimal}$) was obtained for the phosphorus (P) atom charge, 1.14 *e* and 0.857 *e* in the SUN and SVX model compounds, respectively. The ParamChem penalty is relatively high, 31.6 and 83.9 for the P atom in the SUN and SVX model compounds, respectively, indicating that the initial charges should be optimized. With the initial ParamChem charges (the charge of the phosphorous atom of 2.154 *e*), the dipole moment in SUN is just 0.9 D versus 5.2 D computed with QM, while with the optimal charges (the charge of the phosphorous atom of 1.014 *e*) the MM dipole moment improves to 4.2 D. The RMS deviation for ESP also improves from 14.1 kcal·mol^{-1}·Å^{-1} to 1.4 kcal·mol^{-1}·Å^{-1}. The interaction energies were also improved from 0.97 kcal·mol^{-1} to 0.51 kcal·mol^{-1}. Similar improvements were observed for the SVX residue, its results can be found in Tables S571-S573. Overall, these results justify the need to adjust the charge of the phosphoryl group in these compounds.

Finally, to test the impact of the final CHARMM intramolecular geometry on the reproduction of the target water-model compound interactions and dipole moments, they were recomputed with the optimal charge set using the CHARMM optimized geometries. Model compound geometries were optimized using the optimal charges and optimized bonded parameters (see below). The RMS deviation between QM and CHARMM water-compound interaction energies is 0.50 kcal·mol^{-1} very close to 0.46 kcal·mol^{-1} computed using the QM optimized structures. The RMS deviation between QM and CHARMM dipole moments computed with the MM structures is 0.6 D practically identical to 0.6 D computed with the CHARMM optimized structures. The angle between QM and CHARMM dipole moments averaged over all model compounds is 5.0^{o} and 7.7^{o} computed with the MP2/6-31G(d) and CHARMM-optimized geometries, respectively. The RMS deviation between MM and QM ESPs averaged over all molecules is 2.4 and 2.6 kcal·mol^{-1}·Å^{-1} with the QM and CHARMM optimized geometries, respectively. Accordingly, use of the QM gas phase geometries for optimization of the atomic charges yields parameters that are suitable for use with the CHARMM optimized geometries.

Case studies: Optimization of atomic charges for 4-fluorotryptophane (4FW), S-hydroxycystein (CSO), and S-(pyridin-3-ylcarbonyl)-L-cysteine (JJJ)

In this section the charge optimization is exemplified for three amino acids with nonstandard side chains: 4-fluorotryptophane (4FW), S-hydroxycystein (CSO), and S-(pyridin-3-ylcarbonyl)-L-cysteine (JJJ). 4FW is an artificial amino acid, which can be incorporated into proteins to probe thermodynamic and structural properties.[57,58] CSO (also known as sulfenic acid) is an important post-translational modification in proteins, which represents the critical intermediate oxoform in oxidative reactions leading to formation of disulfides, sulfenamides and higher order sulfinic or sulfonic acid species.[59,60] JJJ is a cysteine covalently bound to nicotinaldehyde, an inhibitor of nicotinamidase enzymes[61,62] used to study nicotinamidase function and structure.[63,64] These particular compounds were selected due to their different types of functional groups and, therefore, the presence of different types of interactions with water as well as different polarities. Currently, there are 804, 5, and 2 entries in the PDB for CSO, 4FW and JJJ, respectively. For the charge optimization the appropriate small model compounds were created. The

model compounds include the side chain up to Cα for CSO and JJJ, and up to Cβ for 4FW as presented in Figure 4.

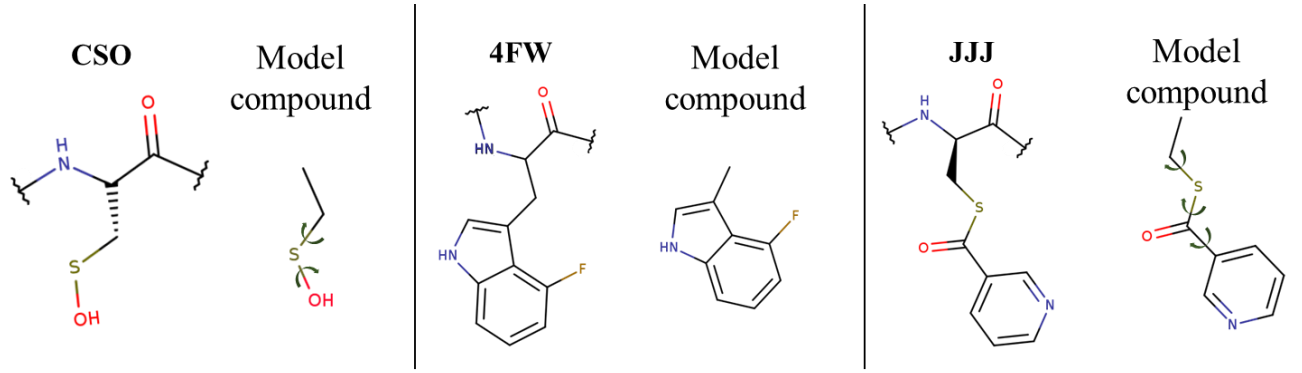

Figure 4. Model compounds used to parametrize 4-fluorotryptophane (4FW), S-hydroxycystein (CSO), and S-(pyridin-3-ylcarbonyl)-L-cysteine (JJJ). The rotatable dihedral angles parametrized in this work are indicated by arrows.

The improvement for selected water interactions is demonstrated in Figure 5. In all cases the dipole moment with the optimized charges is strongly improved relative to the QM dipole moment both in the magnitude and direction. The QM and optimized MM dipole moments for 4FW are 4.5 D and 4.2 D, respectively, while the MM dipole moment with the initial charges is 3.0 D. This improvement was obtained by making the NE1 atom less negative from -0.58 *e* to -0.38 *e*, which also improved the water interaction with the HE1 atom from the absolute error of 0.56 kcal·mol^{-1} to 0.32 kcal·mol^{-1}, with the initial and optimized charges, respectively. However, the charges for the 4FW compound needed only small adjustments, consistent with the small ParamChem penalty for the 4FW compound (the largest penalty of 13.8 is for atom CE3). Larger adjustments of charges were necessary for the CSO model compound. In the CSO model compound, the penalty for atoms OD and SG is very high, 235.7 in both cases, indicating that close analogous groups do not exist in CGenFF. Consistent with this, the water interaction energy computed with the initial charges is 2.95 and 3.60 kcal·mol^{-1} off from the target QM interaction energy, for atoms OD and SG, respectively. The optimized charge for atom OD (-0.56 *e*) is more positive than the initial ParamChem charge (-0.74 *e*), while the charge for SG became more negative: 0.31 *e* against -0.05 *e* for the initial and optimized charge, respectively. The interaction energies with the optimized charges are strongly improved with the absolute deviation from the QM energy of 0.03 and 0.81 kcal·mol^{-1} for atoms OD and SG, respectively. In the JJJ model compound, the dipole moment and angle were improved by making atom O7 less negative from -0.51 *e* to -0.39 *e* and by increasing the negative charge of atom SG from -0.01 *e* to -0.08 *e*, which also helped to improve the water interaction to absolute deviation of 0.11 kcal·mol^{-1} and 0.10 kcal·mol^{-1} for atoms O7 and SG, respectively, compared to the QM results. Atom N1 charge modification from -0.60 *e* to -0.56 *e* lowered the absolute water interaction energy error to 0.06 kcal·mol^{-1} after optimization from initial 0.63 kcal·mol^{-1}.

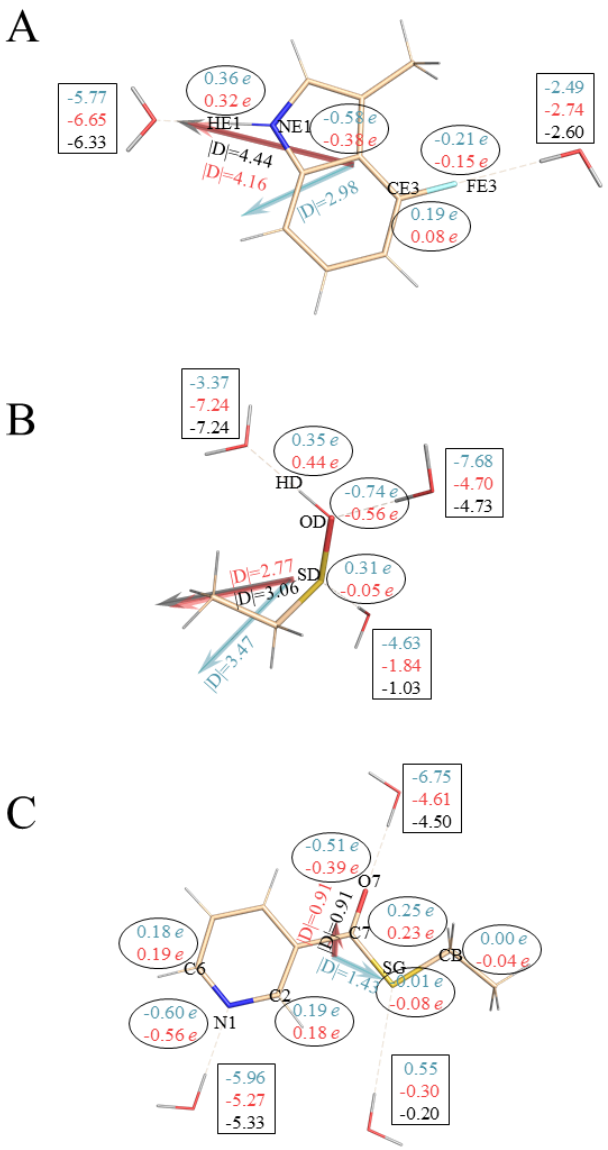

Figure 5. Selected water interactions with model compounds used to parametrize (A) 4FW, (B) CSO, and (C) JJJ. The water-compound interaction energies are given in the rectangular box: MM interaction energies computed with the initial and optimal charges, and QM interaction energies in blue, red and black, respectively. In the oval the ParamChem initial and optimized charges are in blue and red, respectively. The dipole moment computed with the ParamChem charges and optimized charges and the QM dipole moment, are shown as blue, red and black arrows, respectively.

Optimization of bonded terms

All bonded terms including harmonic terms were parametrized based on the reproduction of QM PES. Parameters with the ParamChem penalty > 10 were identified for each amino acid including all accessible protonation/tautomeric forms. A total of 189 model compounds were created to parametrize the bonded terms of 406 nonstandard amino acids. The same terms and associated parameters can be used in several compounds. In such cases, a representative model compound, normally with a fewer number of atoms and with a zero net charge,

was chosen for optimization of a particular bonded term. The optimized parameters were then used for the other model compounds having the same term without further adjustments. Based on this hierarchal approach all model compounds were divided into four groups that contained 132, 24, 14, and 19 model compounds. The first group had all unique parameters, and the subsequent groups have decreasing numbers of free parameters to optimize with the remaining penalty > 10 parameters being optimized in the prior groups. For each term with missing parameters a PES scan was performed. To parametrize all the necessary bonded terms a total of 11194 QM optimizations were performed.

The phase and multiplicity of non-rotatable dihedral angles were taken from the ParamChem guess and were not further varied during optimization, with a few exceptions. In particular, for dihedral angles in conjugated systems the multiplicity was set to two and phase was set to 180º. For improper terms, the equilibrium values were set to zero. The results for bonded term parametrization are presented in this section except for rotatable dihedral angles. The results for empirical and *ab initio* structures and conformation energies are summarized in Table 2. The RMS deviation between the *ab initio* and CHARMM-optimized all Cartesian coordinates averaged over 189 model compounds is 0.18 Å (SD: 0.24 Å) and 0.14 (SD: 0.18 Å) Å for the initial and optimal and parameters, respectively. The values for bonds, valence angles, and torsion angles for the structures optimized with the MM model are in good agreement with the QM values. For bonds, the RMSD between bond distances in QM and MM optimized structures is 0.023 and 0.020 Å, with the initial and optimal parameters, respectively, with the values for valence angles being 1.6º and 1.4º, respectively. For torsions, the RMS deviation is 5.8º and 4.6º with the initial ParamChem and optimized parameters, respectively. Overall, with the optimized bonded parameters, the CHARMM model reproduces the QM geometries very well.

Table 2. Comparison between empirical and *ab initio* optimized geometries for equilibrium structures.

Property	*N* points	MAE	RMSD
		optimal/initial	optimal/initial
[a]RMSD (Å)	189	0.14/0.18	0.18/0.24
bond (Å)	3519	0.015/0.016	0.020/0.023
angle (º)	5968	1.4/1.6	1.9/2.7
dihedral (º)	7133	4.6/5.8	9.6/11.7

[a]RMS deviation between QM and MM optimized equilibrium structures for all atoms.

The RMS deviation between *ab initio* and optimized empirical energies for PES scans is 0.11, 0.31 and 0.43, kcal·mol^{-1} for bond, angle, dihedral and improper angle terms, respectively. For non-rotatable dihedral angles the RMSD is 0.55 kcal·mol^{-1}. The MAE is 0.08, 0.16, 0.30 and 0.23 kcal·mol^{-1} for bond, angle, dihedral and improper angle terms, respectively. Overall, good agreement between QM and MM energies was achieved with correlations between QM and MM relative energies of each data point in the PES of over 90%, except improper angles (86% correlation) as indicated in Table 3. It was found, in agreement with previous studies, that the force

field model well reproduces energies for bonds, angles, but less accurately for dihedral angles.[30] Optimization of parameters improves the agreement between QM and MM energies for all terms. For example, the RMSD for bonds improves from 3.27 to 0.11 kcal·mol^{-1} with the initial and optimal set of parameters, respectively. The significant improvement is explained by the fact that for stiff degrees of freedom even a small deviation in equilibrium values leads to significant deviations in energy. For rotatable dihedrals, the improvement is smaller relative to other degrees of freedom, from 2.29 kcal·mol^{-1} to 0.72 kcal·mol^{-1} with the initial and optimal sets of parameters, respectively.

Table 3. Comparison between empirical and *ab initio* energies of PES scans

Term	[a]*N* terms	[b]*N* points	[c]RMSD	[d]MAE	[e]R
			optimal/initial	optimal/initial	optimal/initial
bond	57	399	0.11/3.27	0.08/1.59	99/17
angle	529	3703	0.31/3.50	0.16/1.39	93/22
stiff dihedral	516	3612	0.55/4.05	0.30/1.07	93/80
rotatable dihedral	212	7844	0.72/2.29	0.43/1.44	96/68
improper angle	24	168	0.43/1.72	0.23/0.78	86/56

[a]Number of terms parametrized in this work; [b]number of PES points used to optimize bonded parameters; [c,d,e]RMS deviation between QM and MM energies, mean absolute error, and linear correlation, R, respectively.

Rotatable dihedrals are degrees of freedom along which the molecule can undergo large structural fluctuations during MD simulations, hence accurate treatment of these dihedral PES is important to describe adequately the conformational space of molecules. Note that the dihedral terms associated with the rotation of the methyl group hydrogens with penalties > 10 were optimized in the present study, although the structural fluctuations due to the rotation of methyl groups are very small due to their being symmetric rotors. Thus, the dihedral terms associated with the rotation of methyl hydrogens were optimized using the method described in the previous section. The parameters of the rotatable dihedrals were determined based on points of PES scans to reproduce the complete rotation of 360º in 10˚ increments, with the exception of methyl groups which were subjected to a 7 point scan due to their symmetry. For all model compounds there were 212 rotatable dihedrals total. The PES points also included the local minimum geometry giving 36 points for each dihedral angle, yielding a total of 7632 QM optimizations to produce the QM target data, which contained 7844 data points.

For rotatable dihedral angles, in contrast to stiff torsion angles, additional Fourier terms (multiplicities or harmonics) were considered and phases were allowed to change from 0 to 180º. In particular, for the residues that have standard C36 parameters for the backbone, three multiplicities (n=1, 2, and 3) were introduced for the torsion terms associated with the rotation around the bond Cα-Cβ (χ1), since χ1 is particularly important for the conformation of the entire side chain. For all other dihedral angles, Fourier series were sought with a minimum number of multiplicities that could fit the energy profiles. However, if a satisfactory agreement was not possible

additional multiplicities were tried. The RMS deviation between QM and CHARMM PES energies for all rotatable dihedral angles and all PES points (7844 total) is 0.72 kcal·mol^{-1}, while MAE is 0.43 kcal·mol^{-1}, demonstrating that the rotatable dihedrals are well reproduced by the force field model. Figure S4 shows the distribution of RMS energy deviation for local minima along PES against the number of molecules. RMS deviation for energy of local minima along PES is lower than 0.5 kcal mol^{-1} for 57% and 80% of the soft dihedral PES scans with the initial and optimized parameters, respectively. As expected, due to the substantial impact of nonbond interactions on their PES the rotatable dihedrals were found the most difficult to fit and the largest RMS deviation with respect to the QM data was observed relative to other harmonic terms and stiff dihedral angles.

Case studies: Optimization of bonded terms for model compounds CSO and JJJ

Here we briefly illustrate the parametrization of bonded terms for CSO and JJJ. Model compounds for bonded term parametrization are shown in Figure 4, and the agreement for geometries is illustrated in Figure 6. In all cases, the MM geometry with the optimized parameters is very close to the QM geometry with the RMS deviation for the Cartesian coordinates of all atoms less than 0.1 Å. The geometries with the initial ParamChem parameters for these two residues are also close to the QM geometries, as can be seen in Figure 6, demonstrating that ParamChem provides a very good guess for these parameters.

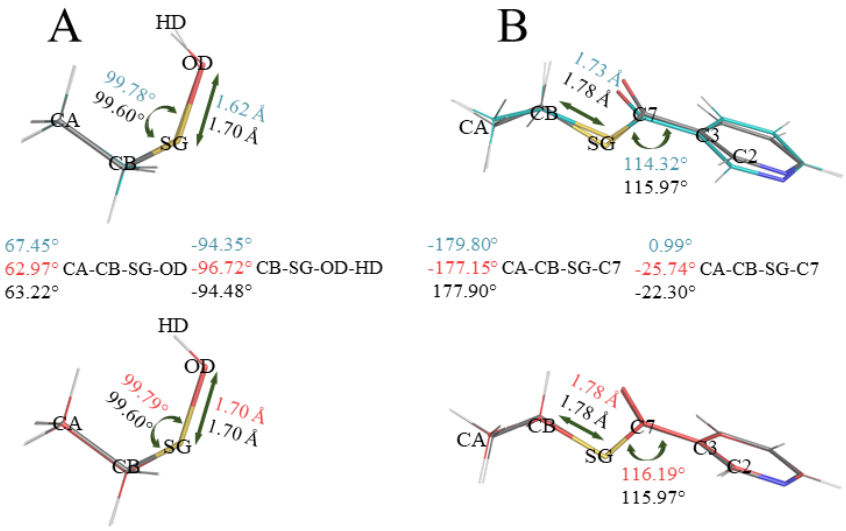

Figure 6. Comparison between QM and MM geometries of (A) CSO and (B) JJJ. The superposition of the QM structure and the structure optimized using the initial and optimal parameters is shown in the upper and bottom panels, respectively. The values for selected degrees of freedom are also given for the QM structure and the structure optimized with the initial and optimized parameters in black, blue and red, respectively.

The agreement between MM and QM energies is demonstrated in Figure 7 and Figure 8 for selected bonded terms in the CSO and JJJ model compounds, respectively, with the initial and optimal sets of bonded parameters. All energies for stiff degrees of freedom are within 2.0 kcal·mol^{-1} of the minimum energy, as described in the Methods section. Overall, the models reproduce well the QM equilibrium conformations of the model compounds as well as QM energies of various deformations along parametrized degrees of freedom. Notable is agreement for rotatable

dihedral angles of CSO model, which involve the rotation of the hydroxyl (C) and the sulfenic (D) groups shown in Figure 7. The position of the local minima and energy barrier heights are well reproduced in both cases. However, the force field model does not reproduce asymmetry of QM PES scans relative to zero degree. In particular, for the rotation the sulfenic group the local minimum at -60º is ~0.4 kcal·mol^{-1} higher in energy relative to the minimum at 60º, while with CHARMM both energy minima have the same energy. This is explained by the fact that in the current CHARMM force field, by convention, the dihedral phases are allowed to be 0º or 180º, so the parameters can be applied for different stereoisomers associated with that dihedral.[30] For the JJJ compound, the deformations along the angle shown in Figure 8 (B) has a non-harmonic energy profile due to the rearrangement of the rotatable dihedral during the PES scan. Similar to the CSO compound, for the JJJ model compound, the force field model well reproduces the PES surfaces associated with the rotatable dihedral angles involving the rotation of pyridine (C) and thiol (D) groups shown in Figure 8. Finally, we note that the positions of wells and barriers for PES in Figure 7 and Figure 8 are in a good agreement for QM energies and MM energies computed with the ParamChem parameters, demonstrating that ParamChem provides a good guess for this molecule.

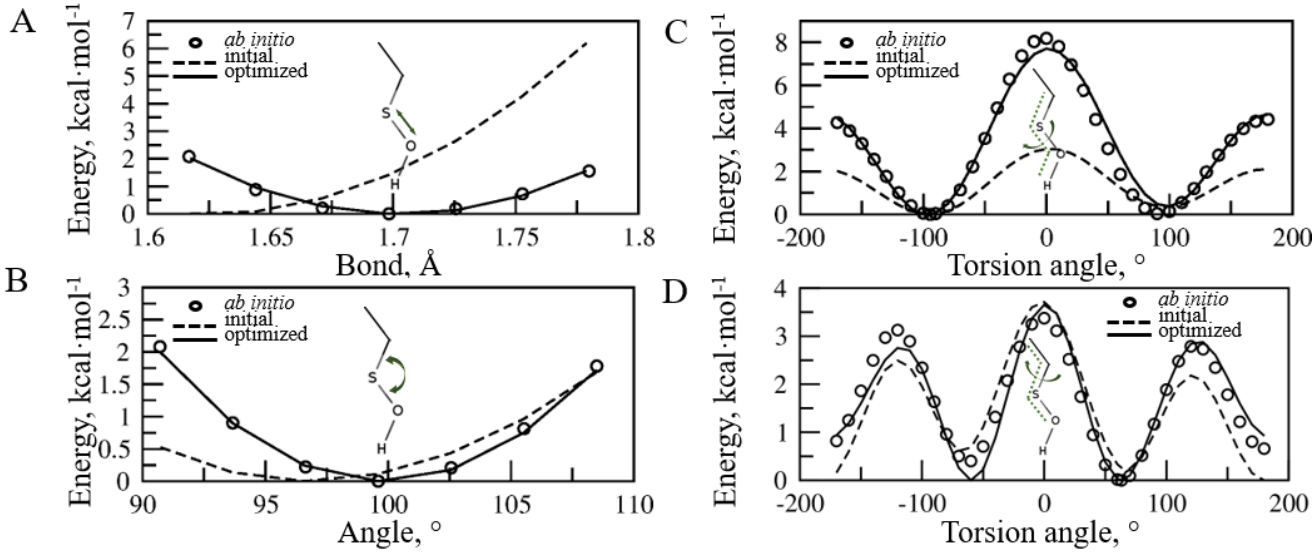

Figure 7. PES scans for selected degrees of freedom in the CSO model compound. The PES scan was performed for a bond (A), for a valence angle (B), and for the dihedral angles associated with the rotation of the hydroxyl group (C) and the sulfenic (D) group, respectively. The arrows and dotted lines indicate the corresponding degree of freedom along which the adiabatic PES scan is performed. The dashed and solid lines show PES energies obtained with the initial and optimal force field parameters, respectively, with the *ab initio* PES indicated by open circles.

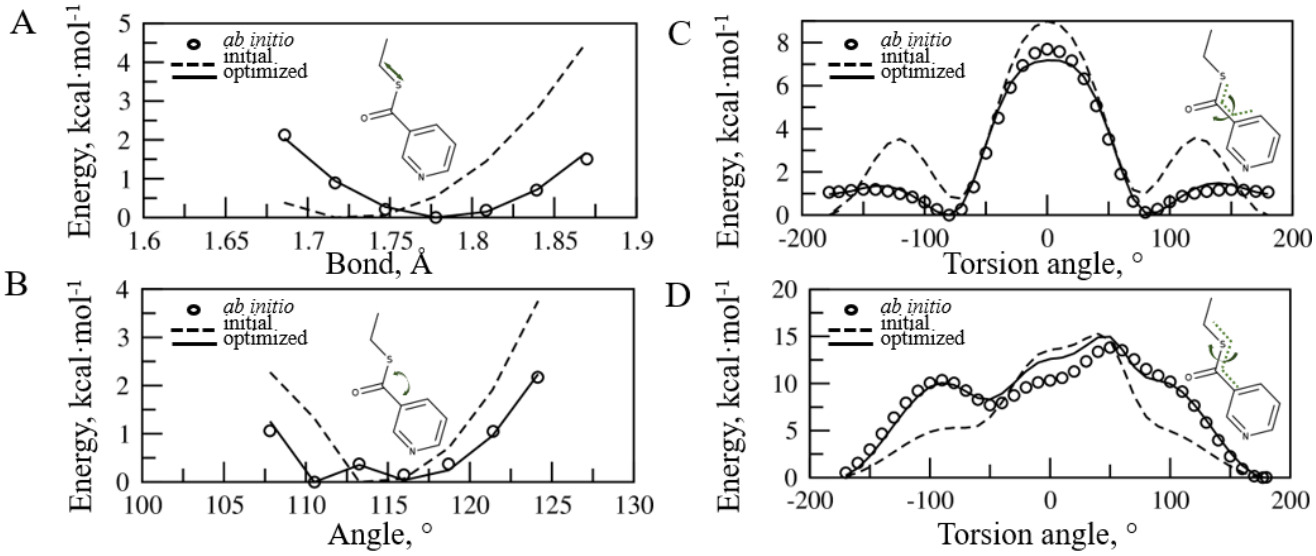

Figure 8. PES scans for selected degrees of freedom in the JJJ model compound. The PES scan was performed for a bond (A), for a valence angle (B), and for the two dihedral angles adjacent to the phenyl ring (C and D). The arrow and dotted lines indicate the corresponding degree of freedom along which the adiabatic PES scan is performed. The dashed and solid lines show PES energies obtained with the initial and optimal force field parameters, respectively, with the *ab initio* PES indicated by open circles.

Molecular Dynamics simulations of protein complexes

To illustrate the quality of the model, MD simulations of proteins containing nonstandard amino acids were performed. Twenty protein structures with a high-to-medium resolution were chosen for MD simulations of 100 ns. The information on the proteins is given in Table S1. Nine systems contained all protein atoms, which were not restrained during MD simulations. A spherical protein model was used with restrained atoms beyond 20 Å for 11 protein systems. Note that our goal was to assess the quality of the force field model for nonstandard amino acids, which should affect primarily the structure and dynamics of the nonstandard amino acid and adjacent residues. The superposition of structures observed in MD simulations on the experimental structures is shown in Figures 9 and 10. For each protein complex 10 snapshots taken every 10 ns from the 100 ns MD simulations were superimposed on the experimental structure based on protein backbone atoms within 10 Å of the nonstandard amino acid. Conformations observed in MD simulations in all protein simulations are very similar to the position of nonstandard amino acids in the crystal structures as shown in Figures 9 and 10. The RMS deviations between simulation and experimental structures are given in Table 4 and S870. The RMS deviation for non-hydrogen atoms within 10 Å of nonstandard amino acids in all MD simulations is in the range of 0.37 Å to 0.99 Å. The RMS deviation for nonstandard amino acids after superimposing on the crystal structure based on the non-hydrogen atoms of the nonstandard amino acid is in the range between 0.11 Å and 0.91 Å; however, the RMS deviation is small for all residues (the mean value for all proteins is 0.40 Å). The largest RMS deviation, 0.91 Å, was observed for carboxymethylated cysteine (residue CCS). CCS106 has a flexible carboxylate group, which rotates during MD simulations starting with the crystal structure 6E5Z.[65] The RMS deviation computed based on non-hydrogen

atoms of CCS without the carboxylate oxygens is much lower 0.49 Å (SD: 0.12 Å), showing that the main contribution to the observed RMS deviation for CCS is due to rotation of the carboxylate moiety. Overall, the RMS deviation for the non-hydrogen atoms of the nonstandard amino acid in all cases is lower than the RMS deviation for unrestrained protein backbone atoms, demonstrating that the model performs as well as the standard CHARMM force field for proteins in protein simulations.

Table 5 and S871 summarize selected non-bond interaction distances. The RMS deviation for distances between non-hydrogen atoms implicated in hydrogen bonds is in the range between 0.0 and 0.59 Å. The largest deviation was observed for residues OCS between N of Lys42 and OD1 of OCS48. In the crystal structure (PDB code: 5IMV),[66] this distance is too short, 2.36 Å, for a hydrogen bond, while in MD simulations the distance between N of Lys42 and OD1 of OCS48 increases to 2.95 (SD: 0.27 Å) Å. The RMS deviation averaged over all distances in Table S871 is 0.18 Å. Thus, important hydrogen bonds between nonstandard amino acids and other protein residues are in very good agreement with the experimental X-ray structures. As an additional test, the rotatable dihedral angles parametrized in this work were further investigated. The torsion angles are given in Table S872. All dihedral angles are well reproduced in MD simulations with the mean absolute deviation from those in the experimental crystal structures of just 5.4º and the RMS deviation of 11.0 º. The largest deviation of 55.5º from the value in the crystal structure (PDB code: 4Y4G)[67] is observed for residue GGB along the dihedral angle defined by atoms Cα, Cβ, Cγ, and Oδ. Further analysis revealed that at the location of atoms Cγ and Oδ there are areas of poor electron density, suggesting that the position of these atoms was not well defined in the crystal model.[67] Indeed, in the crystal structure the distance between atoms CG and NH1 of residue GGB is short, 3.0 Å, so that the distance between their protons of just 1.7 Å creates a repulsion between these groups. In MD simulations this strain is relieved by the rotation around the bond Cβ-Cγ leading to the deviation in the dihedral angle.

Table 4. Root Mean Square (RMSD) deviation in molecular dynamics simulations.

Amino acid	PDB ref.	RMSD (Å) [a]Backbone	[b]Backbone	[c]Residue
MDO	1IYF	0.54 (0.05)/0.49 (0.03)	0.58 (0.08)/0.49 (0.05)	0.33 (0.07)/0.24 (0.07)
4FW	6SZZ	0.72 (0.12)/0.70 (0.12)	0.81 (0.16)/0.78 (0.16)	0.18 (0.06)/0.18 (0.05)
CSO	6Q00	0.80 (0.13)/0.78 (0.12)	0.67 (0.11)/0.62 (0.11)	0.80 (0.12)/0.68 (0.24)

[a]RMSD was computed for unrestrained backbone atoms after superposition on the experimental structure; [b]RMSD was computed based on backbone heavy atoms within 10 Å sphere around the nonstandard amino acid; [c]RMSD was computed for the heavy atoms of the nonstandard amino acid; the numbers are given for MD simulations with the initial and optimized parameters, respectively.

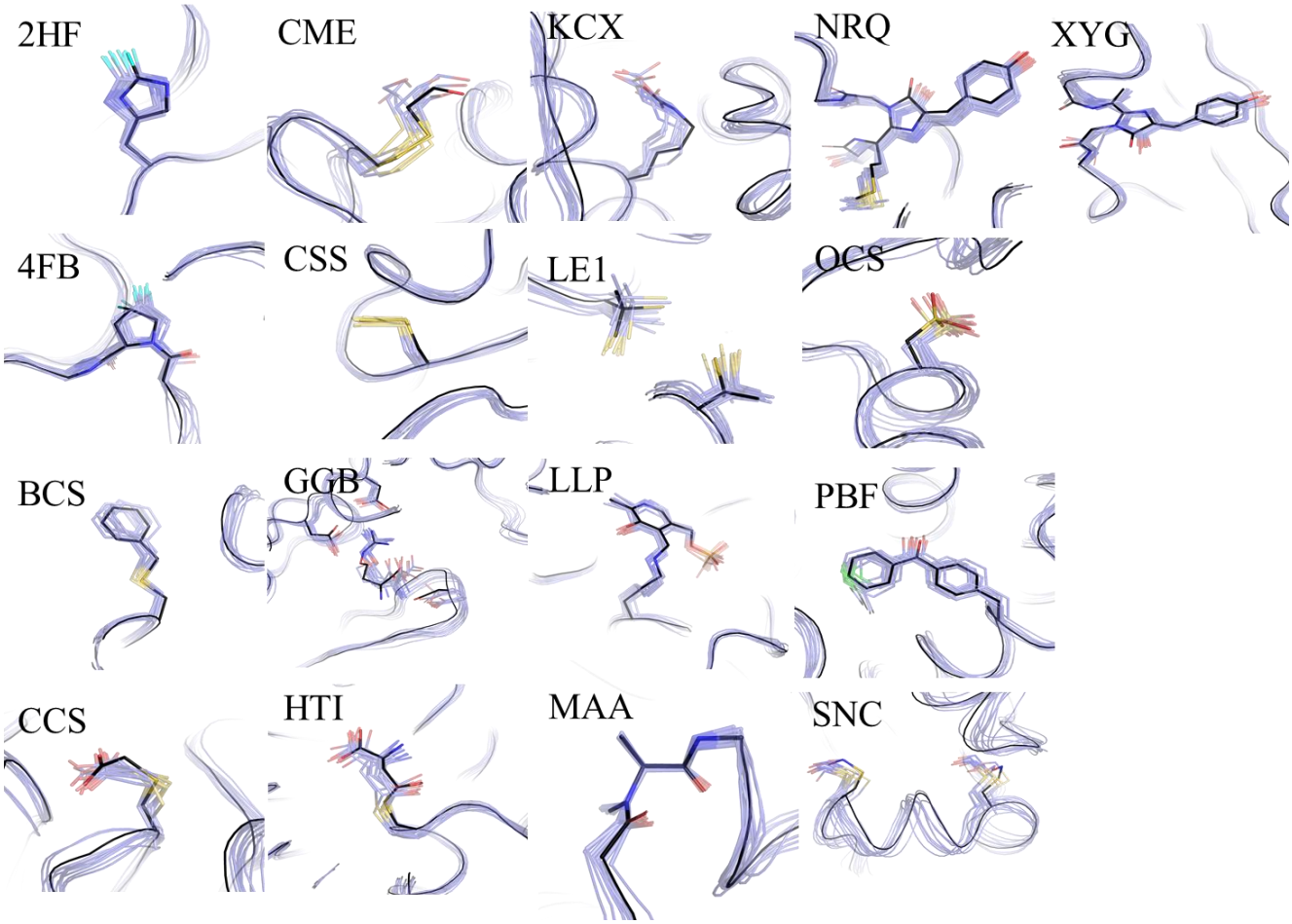

Figure 9. Comparison of structures from MD simulations (in gray) with the experimental structures (in color). Ten snapshots were taken every 10 ns from 100 ns MD simulations and superimposed on the experimental structure using the protein backbone atoms.

In the following, MD simulation results will be presented for three residues CSO, 4FW, and MDO in detail (PDB access codes 1IYF, 6SZZ, and 6Q00, respectively), while the results for simulations of other protein complexes are given in Supplementary Information. Details of the optimization of the parameters associated with CSO and 4FW were presented above. MDO, the 4-methylidene-imidazole-5-one prosthetic group present in phenylalanine-2,3-aminomutase proteins is formed by autocatalytic post-translational modifications of three amino residues (A-S-G) in the polypeptide chain.[68] The RMS deviation, given in Table 4, for the non-hydrogen atoms of the nonstandard amino acid is very low 0.45 (SD of RMSD: 0.03 Å) Å and 0.18 (SD: 0.05 Å) Å, for MDO and 4FW, respectively. For CSO the RMSD for the non-hydrogen atoms is higher 0.68 Å, which is explained by the fact that CSO, in contrast to MDO and 4FW, has two predominant conformations as demonstrated by the analysis of the dihedral angles below. The superposition of the experimental structures for ten snapshots is shown in Figure 10. In all simulations the nonstandard amino acids fluctuate in the vicinity of the experimental position.

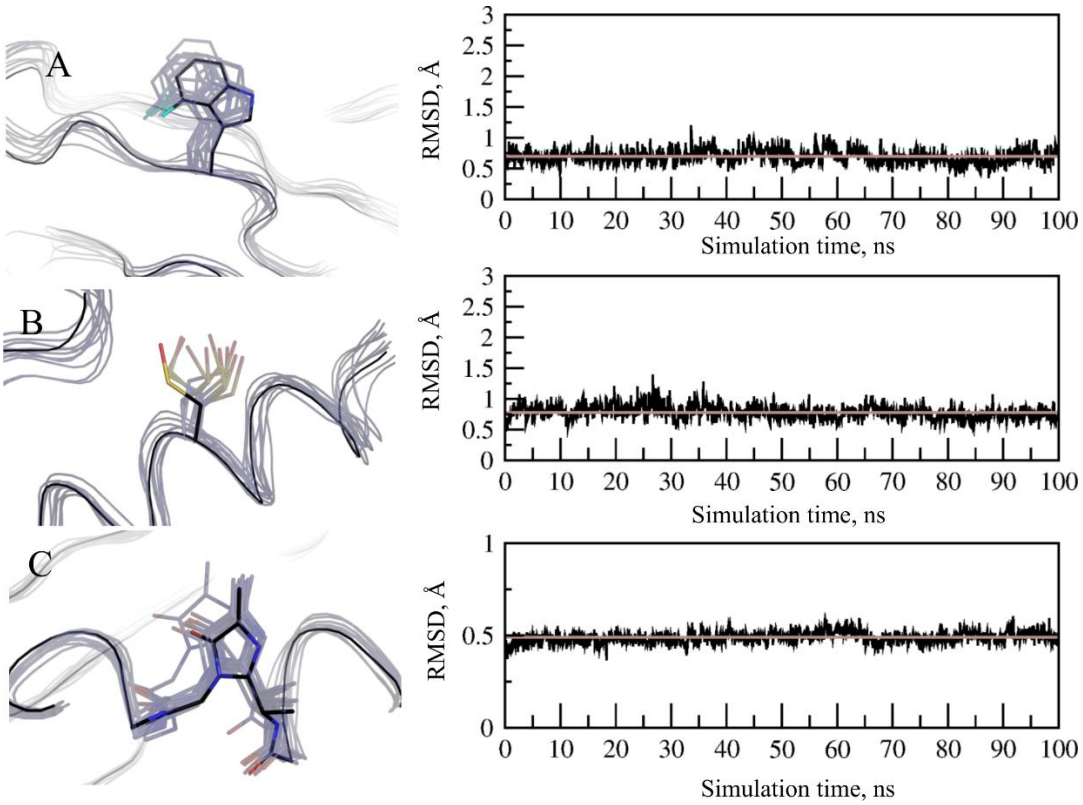

Figure 10. Comparison of structures from MD simulations (in gray) with the experimental structures (in black) for (A) 4FW, (B) CSO, and (C) MDO. (PDB access codes: 6SZZ, 6Q00, and 1IYF, respectively). Ten snapshots were taken every 10 ns from 100 ns MD simulations and superimposed on the experimental structure using the protein backbone atoms. Right panel: RMS deviation for backbone atoms within 10 Å of the nonstandard amino acids; the average RMS deviation is shown in gray.

Table 5. Selected average non-bond distances (Å) in MD simulations of proteins with the nonstandard amino acids.

Residue	[a]Atom pair	X-Ray str.	[b,c] MD simulation	[c]Abs. diff.
MDO	$N2_{MDO66}$-O_{Val61}	2.89	3.04 (0.22)/2.84 (0.12)	0.15/0.05
MDO	O_{MDO66}-$NE2_{Gln94}$	4.19	4.06 (0.65)/3.98 (1.07)	0.13/0.21
MDO	$O2_{MDO66}$-$NH2_{Arg96}$	2.75	2.91 (0.25)/2.88 (0.21)	0.16/0.13
4FW	$FE3_{4FW8}$-N_{Asn10}	3.27	3.45 (0.31)/3.43 (0.28)	0.18/0.16
4FW	$FE3_{4FW8}$-N_{Phe9}	2.94	3.18 (0.22)/3.13 (0.18)	0.24/0.19
CSO	N_{CSO29}-O_{Arg25}	3.05	3.04 (0.21)/3.04 (0.20)	0.01/0.01
CSO	O_{CSO29}-N_{Ala33}	2.93	2.95 (0.19)/2.98 (0.17)	0.02/0.05

[a]Protein atoms (left) are labeled by their amino acid; [b]values in parenthesis are the RMS fluctuations; [c]MD simulations were performed with the initial and optimized parameters, respectively.

Important distances between non-hydrogen atoms are given in Table 5 for MD simulations of MDO, 4FW, and CSO. All average distances observed in the MD simulations are within the RMS fluctuations of the corresponding distances observed in the experimental structures and all within 0.2 Å of the experimental distance. The torsion angles are within the RMS fluctuations from those in the experimental structure for all torsions and

residues. Notable is the agreement for CSO. In the PDB structure 6Q00, two models for the side chain of CSO29 are present with χ1 of 63.6º and 164.3º (model A and B, respectively). Fluctuations around χ1 shown in Figure 11 demonstrate that there are two populated rotamers for CSO in the protein structure with χ1 of 63.2º and 176.1º, both are very close to the experimental values (see also Table 6). Thus, starting from model A, MD simulations with the force field model were able to reproduce both structural models A and B for the CSO side chain. With the initial parameters, the two conformations were also observed in MD simulations (64.6º and 170.1º), however, the conformation with χ1 of ~60º was much less populated, 5.9% and 32.6% with the initial and optimized parameters, respectively. This is due to overestimation of the energy of the conformation at 180 º by 1.2 kcal·mol^{-1} relative to 60º by CHARMM with the initial parameters shown in Figure 7D. MDO, which has a nonstandard backbone group has three rotatable torsion angles and does not have any associated CMAP term. All three torsions can be regarded as dihedral angles in the peptide backbone. The angles observed in MD simulations with MDO are again in very good agreement with those in the X-ray structure, which is important to reproduce the geometry of the entire polypeptide chain. For 4FW, both angles χ1 and χ2 are in excellent agreement with the experimental structure, which is also reflected in the very low RMS deviation between the experimental structure and those observed in the MD simulation. Overall, the model reproduces well the structure of the nonstandard amino acids and their interactions.

Table 6. Rotatable dihedral angles observed in MD simulations and experimental structures. RMS fluctuations are given in parenthesis.

Residue	Dihedral	X-ray	[a]MD	[a]abs. diff
MDO	CB-CA1-C1-N3	-168.7	-160.9 (10.7)/-164.4	7.8/4.3
MDO	C1-N3-CA3-C	101.9	89.2 (9.1)/96.6 (11.1)	12.7/5.3
MDO	N3-CA3-C-N_{V68}	-32.4	-86.5 (22.2)/-50.8 (28.4)	54.1/18.5
4FW	C-CA-CB CG (χ1)	-37.4	-37.7 (7.8)/-37.6 (8.6)	0.3/0.2
4FW	CA-CB-CG-CD1	-89.3	-86.1 (9.0)/-88.6 (8.6)	3.2/0.7
CSO	C-CA-CB-SD (χ1)	63.6/164.3	64.6/170.1/63.2/176.1	1.0/5.8/0.4/11.8
CSO	CA-CB-SD-OD	-149.8	-186.4 (22.3)/-178.1	36.6/28.4

[a]MD simulations were performed with the initial and optimized parameters, respectively.

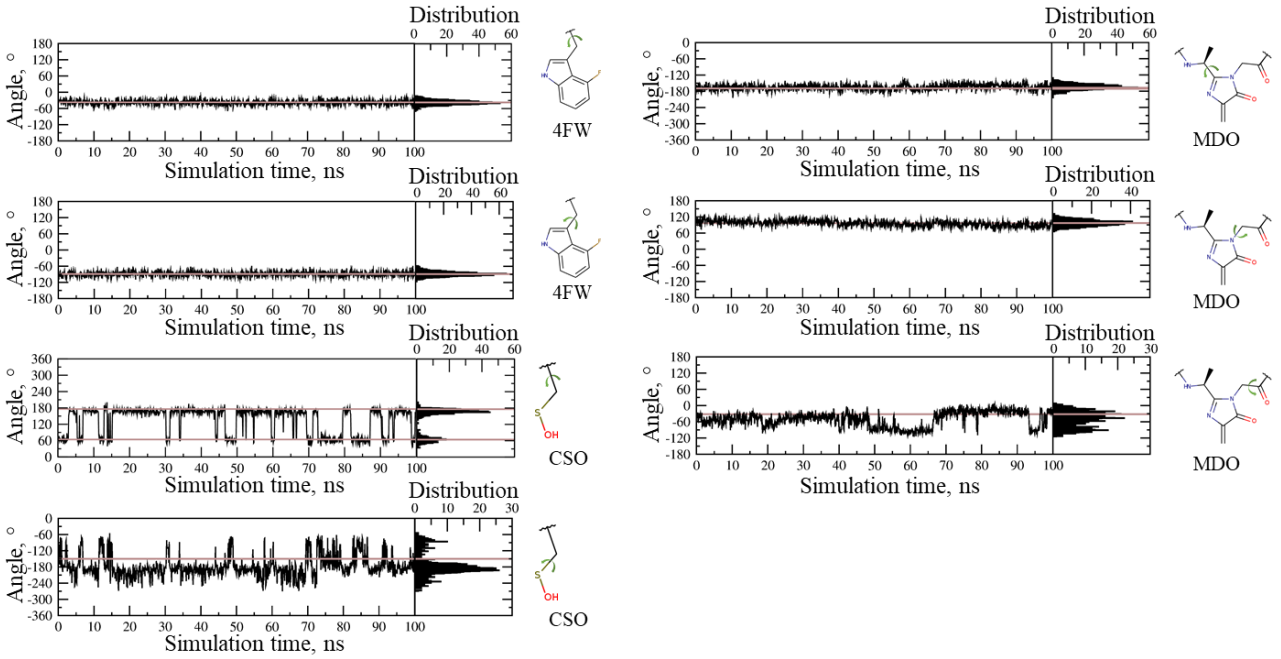

Figure 11. Rotatable dihedral angles in MD simulations of protein complexes with 4FW, CSO, and MDO. (PDB access codes: 6SZZ, 6Q00 and 1IYF, respectively). The dihedral angle is shown by the arrow; the experimental value is shown by the solid gray line; the right panels show the distribution of the dihedral angle in MD simulations.

To test the initial parameters, MD simulations were also performed using the initial CGenFF parameters for CSO, 4FW and MDO. The system setup was identical to the one described above, except the initial CGenFF parameters were used for the modified amino acid. The RMS deviations given Table 4 are systematically larger for the non-hydrogen atoms of the non-standard amino acids, but also for the protein backbone atoms within 10 Å of the modified amino acid. For example for MDO, the RMS deviation is 0.58 Å and 0.49 Å for the backbone atoms, with the initial and optimized parameters, respectively. For the non-hydrogen atoms of the modified amino acids the RMS deviation is also larger with the initial parameters: 0.33 Å vs 0.24 Å with the optimized parameters. Selected dihedral angles given in Table 6 are also systematically better with the optimized parameters. The average non-bond distances, given in Table 5, do not show larger deviations relative to distances in the experimental structures, demonstrating that the ParamChem online server provides a good guess for charges.

Conclusion

The present study represents a systematic development of a force field model for a large set of nonstandard amino acids in the most important protonation states. The parametrization was performed consistent with the standard method used to develop the CHARMM36 additive force field, and thus the model should be compatible with the other components in the CHARMM36 additive force field, including the CHARMM TIP3P water model, the C36 force field for macromolecules and CGenFF for small molecules. The initial guess for both charges and bonded parameters was provided by the ParamChem online server that assigns parameters by analogy from the CGenFF force field. The parameters of the empirical force field were optimized to reproduce QM data and validated against experimental structural data. The charges were adjusted to reproduce interactions of a large number of model compound-water monohydrate complexes, which was important to maintain the balance between interactions of nonstandard amino acids with solvent and other protein residues. In addition, the model reproduces

the scaled magnitude and direction of the *ab initio* dipole moment for neutral compounds as well as the electrostatic potential. Importantly, charge optimization of the neutral species involved systematically overestimating the charges, and thus the dipole moment relative to gas-phase QM data, to introduce implicit electronic polarization corresponding to the condense phase. Including the QM electrostatic potential in the charge optimization, in accord with the previous studies,[37,44] was found useful to obtain a better charge distribution in ionized molecules. Finally, to test that the model well reproduces water interactions with empirical structures of model compounds, probe water interactions were recomputed using the CHARMM optimized structures, demonstrating practically the same level of agreement between force field model results and corresponding QM data.

Special emphasis was given to the quality of all bonded parameters, including soft torsions and stiff harmonic terms, which were adjusted using computationally intensive PES scans. Given the large set of nonstandard amino acids parametrized in this work (406 molecules and their accessible protonation and tautomeric forms) a hierarchical optimization approach, similar to the method used for CGenFF was used for bonded parameters. In this approach, only new parameters that had not been previously available in the force field were optimized, as each new model compound was added to the force field. The order of compounds for bonded parameter optimization was chosen so that the parameters were adjusted in compounds with the minimal possible number of atoms among molecules that share those parameters.

Model validation was based on MD simulations of 20 proteins containing selected nonstandard amino acids. The results demonstrate that the model reproduces very well conformations of nonstandard amino acids in the experimental structures, and in particular rotatable torsions, indicating the quality of both the optimized charges and dihedral parameters. Importantly, the force field model reproduces non-bonded interactions involving the nonstandard amino acids, demonstrating a good balance in the interactions with other components of the system: standard amino acids and water.

The presented parameters represent an extension of the CHARMM36 force field that will allow for reliable molecular simulations of proteins containing nonstandard amino acids. Beyond the parameters for nonstandard amino acids, the parameters developed in this work will be included in the CGenFF force field further expanding its coverage of chemical space. The presented parameters will be incorporated in the program CHARMM[46] and be available from the MacKerell lab web page (https://mackerell.umaryland.edu/) and the CHARMM-GUI (http://www.charmm-gui.org),[15,69] facilitating their utilization in a range of molecular simulation software packages.

Supporting Information

Tables with water-compound interactions; selected distances, root mean square deviations, and rotatable dihedral angles observed in molecular dynamics simulations in protein complexes with nonstandard amino acids; and experimental protein structures used for molecular dynamics simulations.

Acknowledgements

This work was supported by grants ANR-18-CE44-0002 to AA, NIH R01GM138472 to WI, and NIH R35GM131710 to ADM. This work was performed using HPC resources from GENCI-CINES (Grant 2018-A0040710436).

Conflict of Interest

ADM is co-founder and CSO of SilcsBio LLC.

References

(1) Bo¨ck, A.; Forchhammer, K.; Heider, J.; Baron, C. Selenoprotein Synthesis: An Expansion of the Genetic Code. *Trends Biochem. Sci.* **1991**, *16*, 463–467.
(2) Krzycki, J. A. The Path of Lysine to Pyrrolysine. *Curr. Opin. Chem. Biol.* **2013**, *17* (4), 619–625.
(3) Liu, C. C.; Schultz, P. G. Adding New Chemistries to the Genetic Code. *Annu. Rev. Biochem.* **2010**, *79* (1), 413–444.
(4) Magliery, T. J. Unnatural Protein Engineering: Producing Proteins with Unnatural Amino Acids. *Med. Chem. Rev.* **2005**, *2* (4), 303–323.
(5) Khoury, G. A.; Baliban, R. C.; Floudas, C. A. Proteome-Wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot Database. *Sci. Rep.* **2011**, *1* (1), 90.
(6) Mann, M.; Jensen, O. N. Proteomic Analysis of Post-Translational Modifications. *Nat. Biotechnol.* **2003**, *21* (3), 255–261.
(7) Bashir, S.; Harris, G.; Denman, M. A.; Blake, D. R.; Winyard, P. G. Oxidative DNA Damage and Cellular Sensitivity to Oxidative Stress in Human Autoimmune Diseases. *Ann. Rheum. Dis.* **1993**, *52* (9), 659–666.
(8) Gao, W.; Cho, E.; Liu, Y.; Lu, Y. Advances and Challenges in Cell-Free Incorporation of Unnatural Amino Acids Into Proteins. *Front. Pharmacol.* **2019**, *10*, 611.
(9) Almhjell, P. J.; Boville, C. E.; Arnold, F. H. Engineering Enzymes for Noncanonical Amino Acid Synthesis. *Chem. Soc. Rev.* **2018**, *47* (24), 8980–8997.
(10) Hong, S. H.; Kwon, Y.-C.; Jewett, M. C. Non-Standard Amino Acid Incorporation into Proteins Using Escherichia Coli Cell-Free Protein Synthesis. *Front. Chem.* **2014**, *2*, 34.
(11) Sievers, S. A.; Karanicolas, J.; Chang, H. W.; Zhao, A.; Jiang, L.; Zirafi, O.; Stevens, J. T.; Münch, J.; Baker, D.; Eisenberg, D. Structure-Based Design of Non-Natural Amino-Acid Inhibitors of Amyloid Fibril Formation. *Nature* **2011**, *475* (7354), 96–100.
(12) Vanommeslaeghe, K.; MacKerell Jr, A. D. CHARMM Additive and Polarizable Force Fields for Biophysics and Computer-Aided Drug Design. *Biochim. Biophys. Acta* **2015**, *1850* (5), 861–871.
(13) Hagler, A. T. Force Field Development Phase II: Relaxation of Physics-Based Criteria… or Inclusion of More Rigorous Physics into the Representation of Molecular Energetics. *J. Comput. Aided Mol. Des.* **2019**, *33* (2), 205–264.
(14) Sahrmann, P. G.; Donnan, P. H.; Merz, K. M.; Mansoorabadi, S. O.; Goodwin, D. C. MRP.Py: A Parametrizer of Post-Translationally Modified Residues. *J. Chem. Inf. Model.* **2020**, *60* (10), 4424–4428.
(15) Jo, S.; Cheng, X.; Islam, S. M.; Huang, L.; Rui, H.; Zhu, A.; Lee, H. S.; Qi, Y.; Han, W.; Vanommeslaeghe, K.; MacKerell, A. D.; Roux, B.; Im, W. CHARMM-GUI PDB Manipulator for Advanced Modeling and Simulations of Proteins Containing Nonstandard Residues. *Adv. Protein Chem. Struct. Biol.* **2014**, *96*, 235–265.
(16) Khoury, G. A.; Thompson, J. P.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A. Forcefield_PTM: Ab Initio Charge and AMBER Forcefield Parameters for Frequently Occurring Post-Translational Modifications. *J. Chem. Theory Comput.* **2013**, *9* (12), 5653–5674.
(17) Petrov, D.; Margreitter, C.; Grandits, M.; Oostenbrink, C.; Zagrovic, B. A Systematic Framework for Molecular Dynamics Simulations of Protein Post-Translational Modifications. *PLOS Comput. Biol.* **2013**, *9* (7), e1003154.
(18) Margreitter, C.; Petrov, D.; Zagrovic, B. Vienna-PTM Web Server: A Toolkit for MD Simulations of

Protein Post-Translational Modifications. *Nucleic Acids Res.* **2013**, *41*, W422-426.
(19) Reuter, N.; Lin, H.; Thiel, W. Green Fluorescent Proteins: Empirical Force Field for the Neutral and Deprotonated Forms of the Chromophore. Molecular Dynamics Simulations of the Wild Type and S65T Mutant. *J. Phys. Chem. B* **2002**, *106* (24), 6310–6321.
(20) Grauffel, C.; Stote, R. H.; Dejaegere, A. Force Field Parameters for the Simulation of Modified Histone Tails. *J. Comput. Chem.* **2010**, *31* (13), 2434–2451.
(21) Smith, A. K.; Wilkerson, J. W.; Knotts, T. A. Parameterization of Unnatural Amino Acids with Azido and Alkynyl R-Groups for Use in Molecular Simulations. *J. Phys. Chem. A* **2020**, *124* (30), 6246–6253.
(22) Gfeller, D.; Michielin, O.; Zoete, V. Expanding Molecular Modeling and Design Tools to Non-Natural Sidechains. *J. Comput. Chem.* **2012**, *33* (18), 1525–1535.
(23) Gfeller, D.; Michielin, O.; Zoete, V. SwissSidechain: A Molecular and Structural Database of Non-Natural Sidechains. *Nucleic Acids Res.* **2013**, *41* (D1), D327–D332.
(24) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules. *J. Comput. Chem.* **2011**, *32* (11), 2359–2368.
(25) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 490–519.
(26) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.
(27) MacKerell, A. D.; Feig, M.; Brooks, C. L. Improved Treatment of the Protein Backbone in Empirical Force Fields. *J. Am. Chem. Soc.* **2004**, *126* (3), 698–699.
(28) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain X1 and X2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273.
(29) 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* (1), 71–73.
(30) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31* (4), 671–690.
(31) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, first edition.; Clarendon Press, Oxford, 1991.
(32) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52* (12), 3144–3154.
(33) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52* (12), 3155–3168.
(34) Sen, S.; Young, J.; Berrisford, J. M.; Chen, M.; Conroy, M. J.; Dutta, S.; Di Costanzo, L.; Gao, G.; Ghosh, S.; Hudson, B. P.; Igarashi, R.; Kengaku, Y.; Liang, Y.; Peisach, E.; Persikova, I.; Mukhopadhyay, A.; Narayanan, B. C.; Sahni, G.; Sato, J.; Sekharan, M.; Shao, C.; Tan, L.; Zhuravleva, M. A. Small Molecule Annotation for the Protein Data Bank. *Database* **2014**, *2014*.
(35) Westbrook, J. D.; Shao, C.; Feng, Z.; Zhuravleva, M.; Velankar, S.; Young, J. The Chemical Component Dictionary: Complete Descriptions of Constituent Molecules in Experimentally Determined 3D Macromolecules in the Protein Data Bank. *Bioinformatics* **2015**, *31* (8), 1274–1278.
(36) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3* (1), 33.
(37) Huang, L.; Roux, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.* **2013**, *9* (8), 3543–3556.
(38) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 Energy Evaluation by Direct Methods. *Chem. Phys. Lett.* **1988**, *153* (6), 503–506.

(39) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent Molecular Orbital Methods. XX. A Basis Set for Correlated Wave Functions. *J. Chem. Phys.* **1980**, *72* (1), 650–654.
(40) M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, *Gaussian09*, Revision A.01; Gaussian Inc, Wallingford CT, 2009.
(41) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19* (4), 553–566.
(42) Xu, Y.; Vanommeslaeghe, K.; Aleksandrov, A.; MacKerell, A. D.; Nilsson, L. Additive CHARMM Force Field for Naturally Occurring Modified Ribonucleotides. *J. Comput. Chem.* **2016**, *37* (10), 896–912.
(43) Vanommeslaeghe, K.; Yang, M.; MacKerell, A. D. Robustness in the Fitting of Molecular Mechanics Parameters. *J. Comput. Chem.* **2015**, *36* (14), 1083–1101.
(44) Aleksandrov, A. A Molecular Mechanics Model for Flavins. *J. Comput. Chem.* **2019**, *40* (32), 2834–2842.
(45) Press, W. H.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Cambridge University Press: Cambridge, UK ; New York, 2007.
(46) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.
(47) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical PKa Predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525–537.
(48) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* **2004**, *32*, W665-667.
(49) Darden, T. Treatment of Long-Range Forces and Potential. In *Computational biochemistry and biophysics*; Marcel Dekker: New York, NY, 2001.
(50) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.
(51) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.
(52) Aleksandrov, A.; Schuldt, L.; Hinrichs, W.; Simonson, T. Tet Repressor Induction by Tetracycline: A Molecular Dynamics, Continuum Electrostatics, and Crystallographic Study. *J. Mol. Biol.* **2008**, *378* (4), 898–912.
(53) Aleksandrov, A.; Simonson, T. Molecular Dynamics Simulations of the 30S Ribosomal Subunit Reveal a Preferred Tetracycline Binding Site. *J. Am. Chem. Soc.* **2008**, *130* (4), 1114–1115.
(54) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.
(55) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems. *J. Chem. Phys.* **1996**, *105* (5), 1902–1921.
(56) *MarvinSketch*, version 19.19; Chemaxon: Hungary, 2019.
(57) Welte, H.; Zhou, T.; Mihajlenko, X.; Mayans, O.; Kovermann, M. What Does Fluorine Do to a Protein? Thermodynamic, and Highly-Resolved Structural Insights into Fluorine-Labelled Variants of the Cold Shock Protein. *Sci. Rep.* **2020**, *10* (1), 2640.
(58) Budisa, N.; Pal, P. P.; Alefelder, S.; Birle, P.; Krywcun, T.; Rubini, M.; Wenger, W.; Bae, J. H.; Steiner,

T. Probing the Role of Tryptophans in Aequorea Victoria Green Fluorescent Proteins with an Expanded Genetic Code. **2004**, *385* (2), 191–202.
(59) Poole, L. B.; Nelson, K. J. Discovering Mechanisms of Signaling-Mediated Cysteine Oxidation. *Curr. Opin. Chem. Biol.* **2008**, *12* (1), 18–24.
(60) Furdui, C. M.; Poole, L. B. Chemical Approaches to Detect and Analyze Protein Sulfenic Acids. *Mass Spectrom. Rev.* **2014**, *33* (2), 126–146.
(61) French, J. B.; Cen, Y.; Vrablik, T. L.; Xu, P.; Allen, E.; Hanna-Rose, W.; Sauve, A. A. Characterization of Nicotinamidases: Steady-State Kinetic Parameters, Class-Wide Inhibition by Nicotinaldehydes and Catalytic Mechanism. *Biochemistry* **2010**, *49* (49), 10421–10439.
(62) Yan, C.; Sloan, D. L. Purification and Characterization of Nicotinamide Deamidase from Yeast. *J. Biol. Chem.* **1987**, *262* (19), 9082–9087.
(63) French, J. B.; Cen, Y.; Sauve, A. A.; Ealick, S. E. High-Resolution Crystal Structures of Streptococcus Pneumoniae Nicotinamidase with Trapped Intermediates Provide Insights into the Catalytic Mechanism and Inhibition by Aldehydes. *Biochemistry* **2010**, *49* (40), 8803–8812.
(64) Smith, B. C.; Anderson, M. A.; Hoadley, K. A.; Keck, J. L.; Cleland, W. W.; Denu, J. M. Structural and Kinetic Isotope Effect Studies of Nicotinamidase (Pnc1) from Saccharomyces Cerevisiae. *Biochemistry* **2012**, *51* (1), 243–256.
(65) Mussakhmetov, A.; Shumilin, I. A.; Nugmanova, R.; Shabalin, I. G.; Baizhumanov, T.; Toibazar, D.; Khassenov, B.; Minor, W.; Utepbergenov, D. A Transient Post-Translational Modification of Active Site Cysteine Alters Binding Properties of the Parkinsonism Protein DJ-1. *Biochem. Biophys. Res. Commun.* **2018**, *504* (1), 328–333.
(66) Perkins, A.; Parsonage, D.; Nelson, K. J.; Ogba, O. M.; Cheong, P. H.-Y.; Poole, L. B.; Karplus, P. A. Peroxiredoxin Catalysis at Atomic Resolution. *Structure* **2016**, *24* (10), 1668–1678.
(67) Huschmann, F. U.; Linnik, J.; Sparta, K.; Ühlein, M.; Wang, X.; Metz, A.; Schiebel, J.; Heine, A.; Klebe, G.; Weiss, M. S.; Mueller, U. Structures of Endothiapepsin-Fragment Complexes from Crystallographic Fragment Screening Using a Novel, Diverse and Affordable 96-Compound Fragment Library. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2016**, *72* (Pt 5), 346–355.
(68) Barondeau, D. P.; Kassmann, C. J.; Tainer, J. A.; Getzoff, E. D. Understanding GFP Chromophore Biosynthesis: Controlling Backbone Cyclization and Modifying Post-Translational Chemistry,. *Biochemistry* **2005**, *44* (6), 1960–1970.
(69) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L.; MacKerell, A. D.; Klauda, J. B.; Im, W. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **2016**, *12* (1), 405–413.

Table of Contents