Development of force field methods

Anastasia Croitoru

INSTITUT POLYTECHNIQUE DE PARIS

ÉCOLE POLYTECHNIQUE

IP PARIS

NNT : 2022IPPAX093

Thèse de doctorat

Development of Force Field Methods

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Polytechnique

École doctorale n°626 : École Doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Biologie

Thèse présentée et soutenue à Palaiseau, le 14 septembre 2022, par

Anastasia Croitoru

Composition du Jury :

Sophie Sacquin-Mora DR, Institut de Biologie Physico-Chimique (UPR9080)	Présidente
Carine Clavaguera DR, Université de Paris-Sud (UMR8000)	Rapporteuse
Florent Barbault MdC, Université de Paris (UMR7086)	Rapporteur
Bogdan Iorga DR, Institut de Chimie de Substances Naturelles (UPR2301)	Examinateur
Alexey Aleksandrov CR, École Polytechnique (UMR7645)	Directeur de thèse

ACKNOWLEDGEMENTS

I would like to thank all the people that have helped me both in my work and contributed to my personal development.

To begin, I would like to thank my supervisor Alexey ALEXANDROV who in the first place accepted me in the lab first as a Master 2 intern and then as a PhD candidate. I am grateful that he took the chance of advising me during this thesis when I was starting as a biochemist with limited knowledge in bioinformatics and molecular modelling and he successfully guided me to gain expertise of at the time new to me field of scientific research. I learnt to work in a way I was not used to, and it will certainly help me a lot in future projects.

I would also like to thank François HACHE, the director of the Laboratory for Optics & Biosciences (LOB) at Ecole POLYTECHNIQUE for providing me an opportunity to do my research in his lab and together with Nicolas DAVID for supervising the advancement of the PhD and for their help with the administrative part of the thesis.

I would like to acknowledge all the jury members of my thesis. I thank Sophie SACQUIN-MORA for agreeing to be the president of the jury. I thank Carine CLAVAGUERA and Florent BARBAULT for accepting to report my thesis and I thank Bogdan IORGA for agreeing to examine my thesis. I appreciate all of them for their valuable questions and suggestions.

This work would not have been possible without the contributions and exchanges with our collaborators. I thank Muriel GONDRY for all the enriching discussions around cyclodipeptide synthases and Morgan BABIN for the physical experiments as some of my thesis work is based on their experimental results. I am extremely grateful to Alexander MACKERELL for his expertise and all the exchanges during CHARMM force field development. I thank Anmol KUMAR for conducting additional computational calculations. I am also grateful to Suliman SHARIF for the help with selecting the ZINC custom database and all his precious advice regarding the use of Python and its application to computational chemistry. I also thank Wonpil IM, Jumin LEE and Sang-Jun PARK for making available from CHARMM-GUI webserver the force field developed during this PhD.

During this PhD, I discovered my passion, for teaching for which I would like to thank Roxane LESTINI as she trained me and helped me setting up the experiments for the practical classes. I am forever grateful for her availability, kindness, and pedagogy.

I also thank Ernan ROITMAN and Jean-Christophe LAMBRY for their help in setting up the computer machines and for helping me solve all the technical problems that came along. I also enjoyed participating at the gardening moments they organised outside our lab and for sharing conscious principles of living with respect towards nature.

I want to express my thanks to Jean COGNET who was one my teachers during my Master studies. I learned about force fields and molecular modelling for the first time during his classes and right away I knew what I wanted to study in the next years. I am very grateful for all the discussions we had and for all his patience in answering all my questions regarding his course.

I am also grateful to all the members of LOB for their kindness and warm welcome, especially to PhD students Margaux SCHMELTZ, Clothilde RAOUX and Yoann COLLIEN for the great time during our breaks. I thank Maëlle VILBERT, Sengbin LIM, Hugo BLANC, Júlia FERRER ORTAS and Robin KUHNER for their friendship, ambiance in the lab or during breaks and for sharing together the experience of a PhD.

I am particularly fortunate to have met Seongbin LIM, a friend and exceptional personality who has proven that there are no limits to kindness. I thank him for making the lab feel like home.

I am grateful for Corentin FOSTIER's and Eirini CHRYSANTHOU's friendships. Both, they were an inspiration and motivation for me to get over all difficulties and to appreciate all the accomplishments.

I want to express my thanks to Mieke VEUCHELEN and Noël HENDRICKX for encouraging at a crucial point in my life when I decided to come to France. I don't know how things would have evolved without their friendship.

A special thanks goes to Ada, who helped me deal with all the stress or anxiety. She taught me to be more active, accept with ease all types of situations and to do it with an open mind.

Finally, I would like to thank my family. I thank my mother and grandparents for supporting my choices, for encouraging me to study and for their love. I am grateful to my brother Florin for his positivity, interest in my research and for being able to always make me smile, even in the hardest moments. I especially want to thank Nicu, my husband, who encouraged me, cared for me, and shared with me all the ups and downs of this thesis. I treasure every moment spent with you and I express my gratitude for all your patience, compassion, and goodwill.

SUMMARY

Drug discovery and development are very time and resources consuming processes, which can be significantly facilitated by computer-aided drug design methods. Among such methods force field-based are arguably the most used. The goal of this PhD was to develop the force field (FF) model for a large number of biologically important molecules.

In the first part, I focused on extending the CHARMM force field to a large set of 333 nonstandard amino acids. These nonstandard amino acids are frequent in the protein structures available in the protein data bank. These are biologically important molecules produced as a result of post-translational modifications (PTMs) in the cell, but also can be synthetized and incorporated in labs. For parametrization, amino acids with nonstandard sidechains as well as amino acids with modified backbone groups were considered. Amino acids were parametrized for the most important protonation states at physiological pH and, for some more common residues, in both D- and L-stereoisomers. Both inter- and intramolecular terms were parametrized targeting quantum mechanics (QM) data. Validation was performed by molecular dynamics simulations of 20 protein systems.

During my PhD, I also worked on the development of a force field model for the phenylalanyl group covalently linked to $tRNA^{Phe}$. This model was used to predict the structure of an enzyme of the novel cyclodipeptide synthases family bound to phenylated $tRNA^{Phe}$. In collaboration with an experimental group, I showed that the proposed model is compatible with experimental data.

Another part of my PhD was dedicated to improving the force field development method, where I developed and tested a new method for bond and valence angle terms parametrization to improve transferability and robustness of developed parameters. The novelty of the method is that it allows explicitly structural deviations between QM and CHARMM structures during optimization. The results demonstrate that without any need for additional restraints the new method produces robust and transferable force field parameters. The new method also improves the agreement for the QM normal modes for all molecules in the set. Thus, the new method will allow parametrization of molecules under the structural

deviation, common for force fields for small molecules, producing robust and transferable parameters.

In the final part of the project, I am performing a large-scale parametrization of drug-like molecules. Around 300,000 ligands were selected from the ZINC20 database to cover a broad region of the chemical space. The selection criteria accounted for the drug-likeness and chemical diversity of molecules, which are not parametrized in the current CHARMM force field. Based on the sorting method to find molecules containing common atom groups, 7000 molecules were subjected to the force field development. A special attention was given to the optimization of non-planar rings, which can exist in different puckering states.

Overall, the current PhD project represents a significant step forward in extending the CHARMM FF to a wide variety of chemical entities. The FF model for both, nonstandard amino acids and the ligand library, apart from structure/function studies can be used for virtual screening studies. To make available for the scientific community, the FFs developed in this work are included into the standard CHARMM package of FF.

RÉSUMÉ

La découverte et le développement de médicaments sont des processus très coûteux en termes de temps et de ressources, qui peuvent être considérablement facilités par des méthodes de conception de médicaments assistées par l'ordinateur. Parmi ces méthodes, celles basées sur les champs de force sont sans doute les plus utilisées. L'objectif de cette thèse était de développer un modèle de champ de force pour un grand nombre de molécules biologiquement importantes.

Dans la première partie, je me suis concentrée sur l'extension du champ de force CHARMM à un large ensemble de 333 acides aminés non standard. Ces acides aminés non standard sont fréquents dans les structures protéiques disponibles dans la banque de données des protéines. Il s'agit de molécules biologiquement importantes produites à la suite de modifications post-traductionnelles dans la cellule, mais elles peuvent également être synthétisées et incorporées dans les laboratoires. Pour la paramétrisation, les acides aminés avec des chaînes latérales non standard ainsi que les acides aminés avec des groupes de squelette carboné modifiés ont été considérés. Les acides aminés ont été paramétrés pour les états de protonation les plus importants au pH physiologique et, pour certains résidus plus communs, dans les formes stéréoisomères D et L. Les termes inter- et intramoléculaires ont été paramétrés en fonction des données de la mécanique quantique (MQ). La validation a été effectuée par des simulations de dynamique moléculaire de 20 systèmes protéiques chacun contenant un acide aminé non standard différent.

Au cours de mon doctorat, j'ai également travaillé à l'élaboration d'un modèle de champ de force pour le groupement phénylalanyle lié de façon covalente à l'ARNt^{Phe}. Ce modèle a été utilisé pour prédire la structure l'ARNt^{Phe} phénylalanylé liée à AlbC, une enzyme représentante d'une nouvelle famille enzymatique appelée synthétases de cyclodipeptides. Le modèle a été obtenu d'abord utilisant l'amarrage moléculaire rigide et ensuite raffiné par des longues simulations de dynamique moléculaire au cours desquelles le complexe de l'ARNt et AlbC est maintenu avec une interaction stable entre les deux partenaires moléculaires. En collaboration avec une équipe d'expérimentalistes, j'ai montré que le modèle théorique proposé est compatible avec les résultats expérimentaux.

Pour améliorer la méthode de développement du champ de force, j'ai développé et testé une nouvelle méthode de paramétrage des termes de liaison et d'angle de valence afin d'améliorer la transférabilité et la robustesse des paramètres développés. La nouveauté de la méthode consiste à en permettre explicitement les déviations structurelles entre les structures MQ et CHARMM pendant l'optimisation du champ de force. Les résultats démontrent que sans aucun besoin de contraintes supplémentaires, la nouvelle méthode produit des paramètres de champ de force robustes et transférables. La nouvelle méthode améliore également l'accord CHARMM et MQ pour les modes normaux pour toutes les molécules de l'ensemble utilisé. Ainsi, la nouvelle méthode permettra la paramétrisation des molécules sous une déviation structurelle, courante pour les champs de force des petites molécules, produisant des paramètres robustes et transférables.

Dans la dernière partie du projet, je réalise une paramétrisation à grande échelle de molécules de type médicament. Environ 300.000 ligands ont été sélectionnés dans la base de données ZINC20 pour couvrir une large région de l'espace chimique. Les critères de sélection tiennent compte de la similitude avec les médicaments et de la diversité chimique des molécules, qui ne sont pas paramétrées dans le champ de force CHARMM actuel. En se basant sur la méthode de tri pour trouver les molécules contenant des groupes d'atomes en commun, 7000 molécules ont été soumises au développement du champ de force. Une attention particulière a été accordée à l'optimisation des cycles non-planaires, qui peuvent exister dans différents états de conformation.

Dans l'ensemble, le projet de thèse actuel représente une avancée significative dans l'extension du champ de force CHARMM à une grande variété d'entités chimiques. Le modèle de champ de force pour les acides aminés non standard et la bibliothèque de ligands, en plus des études structure/fonction, peut être utilisé pour des études de criblage virtuel. Afin de les mettre à la disposition de la communauté scientifique, les champs de force développés dans ce travail sont inclus dans le paquet standard du champ de force CHARMM.

CONTENTS

Chapter 1
INTRODUCTION

In silico studies have been become routine for biomedical and material research nowadays, thanks to advances in algorithms as well as hardware. They allow reducing the time and cost for experimental studies, and in particular, for expensive drug design campaigns. Among *in silico* tools, molecular mechanics (MM) based methods are arguably the most popular computational techniques for studies of biomolecular systems owing the system size and timescales that can be accessed. The highest accuracy of modelling is obtained by quantum mechanics (QM) methods that find solutions to the Schrodinger equation. However, such methods are still computationally expensive and cannot be applied to simulate biological processes occurring on long timescales, such ligand-protein binding or protein conformational changes. MM methods approximate QM Hamiltonian using an empirical function, also called a *force field* (FF) that defines the energies and forces acting on the molecular system.

The notion of *a force field* includes an empirical energy function, a set of associated empirical parameters that need to be fitted, and a parametrization strategy. The parameter set and potential energy function combined allow to calculate the energy and forces as a function of coordinates of the particles in the system. Using atomic forces, it is possible to solve numerically Newton's equations of motion, by molecular dynamics (MD) simulation techniques. The quality of such simulations depends dramatically on the accuracy of the underlying FF model, and in particular FF parameters. The major requirement for performing such MD simulations is the existence of MM force field parameters. While FF models exist for a set of standard molecules such as standard amino acids, nucleic acids, lipids, carbohydrates, and a limited number of drug-like small compounds, in general, a FF model is not available for every molecule in the infinite chemical space.

The main goal of this thesis was to extend the CHARMM force field family to cover a larger domain of the chemical space. The first objective was to generate high-quality (i.e. specifically optimized) FF parameters for a large set of frequent nonstandard amino acids. The second objective was to develop a new improved

method for the FF development for bonded terms. The final goal of the thesis was to perform large-scale FF parametrization for small drug-like molecules from a library, which can be used in *in silico* drug design endeavors. In what follows, I will give a brief introduction to FFs with an accent on the additive version of the CHARMM family of FFs.

FORCE FIELDS (FFS)

Force fields can be classified by the smallest entity, which degrees of freedom are explicitly present in the model: electrons, atoms, groups of atoms, tissues and so on. The *coarse-grained* models like SIRAH[1], VAMM[2] or MARTINI[3] represent groups of atoms as a single entity, called a "bead". Often, the beads have a dipole moment or higher order electric moments, and complicated functions are used for bonded interactions and hydrogen bonds to restore physical properties lost by atoms in the bead model. Coarse-grained models provide benefits for simulating massive biomolecular systems and to study functional processes occurring on larger time scales,[4] however, in the context of computer aided drug design a more precise all-atom force field is needed. Similar to the current coarse-grained models, the initial versions of OPLS, CHARMM, AMBER used the *united-atom* concept,[5–7] while the GROMOS force field still inherits this approximation.[8] The *united-atom* model does not include explicitly nonpolar hydrogens, instead parent atom parameters are optimized to account implicitly for steric effects due to missing hydrogens. By contrast, polar hydrogens are treated explicitly to model hydrogen bonds and any other polar interactions.[9]

All atom FFs typically have two contributions, one from intermolecular and one from intramolecular interactions. Both types of interactions are present even in a small molecule in vacuum. Generally, the intermolecular part is due to electrostatic and van der Waals (vdW) interactions. Intramolecular interactions are related to the covalent structure of the molecule. Depending on the type of terms included to the potential energy function, FFs can be divided into three classes. In the Class I force fields, deformations of bonds and angles are described by simple harmonic oscillator. Thus, the magnitude of the restoring force is assumed proportional to the displacement from the equilibrium position. However, bond stretching and angle bending are harmonic only near the equilibrium values, in the limit of small vibrations, and higher-order terms may be required for a more accurate description of molecular motions. Force fields in the class II include anharmonic cubic and higher

order contributions to the potential energy for bonds and angles, and can also include cross-terms describing the coupling between bonds, angles and dihedral angles resulting in a more accurate reproduction of experimental bond and angle vibrations. However, FF models of this class (such as MMFF94[10], UFF[11]) require more parameters to be fitted in comparison to class I.

Class I and II employ the Coulomb law to model electrostatic interactions with fixed point (partial) charges centered on atoms. This model is referred to as "additive", as electrostatic interactions between atom pairs are pairwise in nature, i.e. not affected by the presence of other charges in the system. The main disadvantage of additive FFs is that it does not account explicitly for polarization. With partial charges being fixed during simulations, the induced polarization is treated implicitly in a mean-field way. By contrast, the electronic density of the molecule is not static and can adjust in response to the external electric field, affecting interactions with other molecules, but also with itself. To solve this problem, Class III FFs have been developed that treat explicitly electronic degrees of freedom. Examples of Class III FF are AMBER ff02[12–15], AMOEBA[16–21], CHARMM-FQ[22–26], and CHARMM Drude[27–53].

In this work, I focus on the additive version of FFs for its computer efficiency, appreciated in the field of CADD. Typically, for bio-systems molecular dynamics simulations are performed close to room temperature and energy of bond and angles vibrations stay low enough to be modelled by harmonic terms[54] of Class I FF (even at temperatures specific to thermophiles). Although the Class I FF lacks the explicit treatment of electronic polarizability, a common strategy (as for the additive CHARMM FF) is to include it implicitly by overestimating the corresponding gas-phase molecular dipole moment by ~20-30%.[55] This strategy is based on the fact that dipole moments of molecules in condensed phases are normally larger than those in the gas phase. As consequence, the CHARMM fixed charge model has shown good agreement with condensed phase properties, including experimental molecular volumes and enthalpies of vaporization.[9,56–63]

In all-atom additive force fields, each atom is represented by a point in space with a mass, a partial charge, van der Waals parameters, and connected to other atomic points by bonded terms. The dynamics of such systems of atoms system can be described by Newton's equations of motion. The classical molecular force fields

share essentially the same empirical expression for the potential energy proposed by Levitt and Lifson in 1969:[64]

$$E_{pot}(\vec{r}) = E_{bond}(\vec{r}) + E_{angle}(\vec{r}) + E_{torsion}(\vec{r}) + E_{improper}(\vec{r}) + E_{elec}(\vec{r}) + E_{vdW}(\vec{r})$$

(Eq 1.1)

$\vec{r}$ are the coordinates of atoms. E_{bond}, E_{angle}, $E_{torsion}$, and $E_{improper}$ are bonded terms due to bond stretching, angle bending, dihedral angle bending, and improper dihedral angle bending. The non-bonded terms E_{elec} and E_{vdW} describe the Coulomb (electrostatic) and van der Waals interactions between atoms not directly connected via covalent bonds and bond angles (1-4 atoms in the CHARMM FF). A graphical representation of the terms in the potential energy function are shown in Figure 1.1.

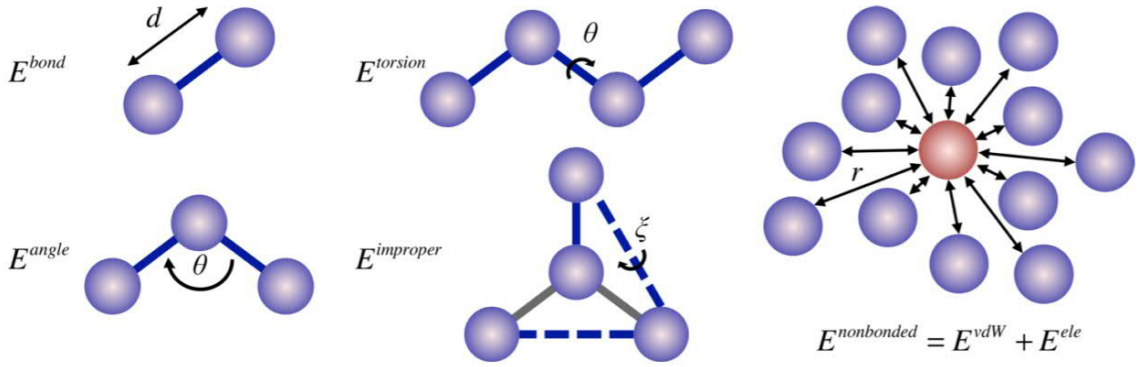


Figure 1.1. Schematic illustration of the terms in a classical fixed-charge force field. Bond stretching (E^{bond}), angle bending (E^{angle}), dihedral angle torsion ($E^{torsion}$), and improper angle bending ($E^{improper}$) as well as van der Waals (E^{vdW}) and electrostatic (E^{ele}) interactions are shown.[65]

The major families of all-atom additive force fields are Amber (Assisted Model Building with Energy Refinement)[66,67], OPLS (Optimized Potentials for Liquid Simulations),[58,68] and CHARMM (Chemistry at Harvard Macromolecular Mechanics).[69–71] These force fields employ atom types to define bonded and non-bonded parameters. For a given molecule, from atom types and connectivity information it is sufficient to determine all bonded parameters.[65] The notion of atom type is related to the fundamental assumption behind FFs, called transferability that the same group of atoms behave similarly in different chemical species. For example, the angle between protons of the methyl group as well as vibrations along this angle are very similar in different molecules. The transferability feature allows developing FFs for novel molecules in a hierarchical manner, where fitted parameters for novel molecules are included to the force field to serve for other molecules not present yet in the transferable force field.[72] There are fewer atom types in a FF in comparison to

a number of molecules that can be represented by the force field model. Generally, the number of atom types is increasing with the number of molecules, requiring more "specialized" atom types. For example, the current CHARM36 force field for proteins contains 53 atom types, while the CHARMM FF for small molecules (CGenFF) has already 163 different atom types.[65] Including additional atom types for the same set of molecules generally improves agreement with QM data allowing capturing more subtleties, but also leads to more parameters to fit. Before going into details of force fields for biomolecules, the MM models for water will be described, as the water molecule was historically one of the first molecule described by MM.

Application of FFs in the context of biological systems requires to include the effect of solvent as most of protein biological function take place in an aqueous environment. Aqueous solvent plays a determining role in the majority of biological processes, including protein folding and ligand-protein binding. For example, upon formation of ligand-protein complexes the solvent content of the binding site changes significantly by replacement of the polar aqueous solvent by polar and non-polar groups of the ligand. In general, the ligand-protein binding represents a fine balance between ligand-protein, protein-solvent, and ligand-solvent interactions. Another example is protein folding, where an amino acid in the polypeptide chain changes its solvent environment to the folded-protein environment, which can be significantly non-polar. The fine balance of interactions between solutes-solutes, solutes-solvent and solvent-solvent needs to be reproduced by FF in simulations. To model these interactions, the most accurate approach is to use atomistic models (explicit solvent), which provides the highest level of detail of biomolecular systems. However, implicit solvent models, though not developed or used in the current work, also should be mentioned in this context, as a significant amount of effort has been invested in the development of approximate schemes.[73]

Creating models for water is a complicated task due to water's unusual physical and chemical properties. Hundreds of potential models for explicit water have been proposed aiming to reproduce a specific nature of the molecule, with the first models dating to the first half of the twentieth century. Similar to force fields for biomolecules, water models can be classified by the treatment of electrostatic interactions, i.e. polarizable and with fixed charges. The fixed charge models can be further classified by whether the water geometry is flexible or rigid, by the number of sites that need to describe vdW and electrostatic interactions.

Figure 1.2 gives a schematic representation of the different water models available[74] and also outlines a historical aspect on the development of water FF models. One of first water models was proposed by Mecke and Baumann in *circa* 1930[75]. The model consists of two positive charges on each hydrogen and a negative charge on the HOH bisector, while the vdW interaction site was centred on the oxygen atom. A 4-sites water model was proposed later by Bernal and Fowler,[75] parametrized to reproduce experimental data (structure of ice, dipole moment, diffraction pattern of liquid water). In this model, each hydrogen had a positive charge, and two negative charges were placed "behind" the oxygen atom to form a tetrahedral geometry. A different model was proposed by Rowllinson (Figure 1.2c) similar to Bernal and Fowler, also with four electrostatic sites: in addition to the two positive charges on the hydrogen atoms, two negative charges were placed near the oxygen atom[76]. This model better accounts for the lattice energy and intermolecular spacing in ice, including the effects of polarization and induced dipole. Ben-Naim and Stillinger (BNS) model, proposed in 1972[77] and represented in Figure 1.2d, was used to perform the first molecular dynamics study of water. This model, similar to Rowllinson's model, has four-point charges, but the two negative charges near the oxygen atom were placed by the authors in a tetrahedral geometry. The Transferable Intermolecular Potential, with three centres (TIP3P) was developed by Jorgensen[78] in 1981 yr. with the purpose of a simple and efficient water potential that can be transferable to solute-solvent systems and requires less computer resources (Figure 1.2b) in contrast to 4-site models. This model is based on three sites, two positive charges on each hydrogen and one negative charge on the oxygen atom, with the only vdW site placed on the oxygen centre, again to reduce the computational cost. The parameters were based on gas-phase dimers and pure liquids structures and energies. Analogous to TIP3P, Berendsen and co-workers derived the Simple Point Charge (SPC) water potential,[79] the difference being that the parameters were optimized to reproduce experimental potential energies and pressure of liquid water.

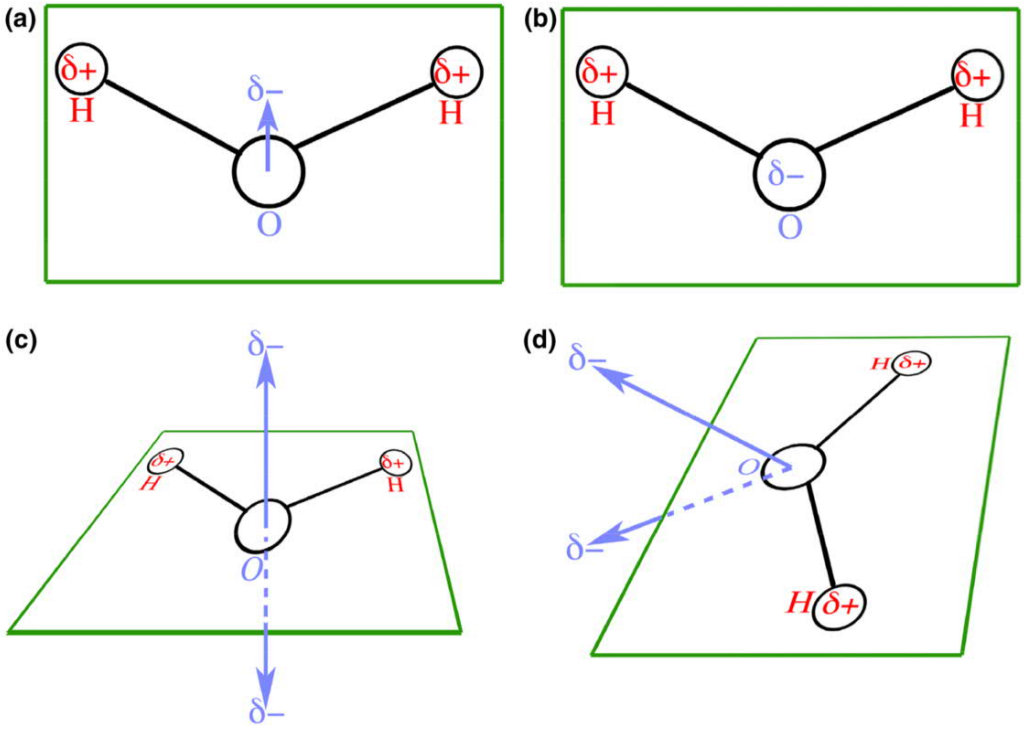


Figure 1.2. Schematic representation of different water models. a) TIPS2, TIP4P; b) SPC and TIP3P; c) Rowlinson; d) Ben-Naim and Stillinger (BNS) and Stillinger (ST2). The plane of the molecule is indicated by a green parallelogram, oxygens and hydrogens are in blue and red, respectively. δ is the absolute value of partial charges.[74]

As it was shown,[80] Bernal and Fowler potential does not reproduce well the bulk properties of water, such as density and heat of vaporization. Later models, such as TIPS2 and TIP4P follow the Bernal and Fowler's idea by adding one dummy atom near of the oxygen atom along the bisector of the HOH angle and aim to reduce these inconsistencies. These models, shown in Figure 1.2a, are able to reproduce with a better accuracy the bulk properties.[80]

Historically, the water model used with the first version of CHARMM FF was ST2. ST2[81] potential was developed by Raman and Stillinger by slightly modifying the BNS model. The main difference between BNS and ST2 is the distance from the oxygen atom at which the negative charges are placed. Modern simulations with the additive CHARMM force field mostly use a modified version of the TIP3P water model, (ambiguously) called TIP3P (modified version), which is also a water model of choice for simulations with the Amber force field. The modified TIP3P is similar to the unmodified TIP3P but contains additional vdW interaction sites apart from the oxygen site, which are placed at hydrogens. In particular, TIP3P water model appears, while being computer efficient, to reproduce reasonably well different physical properties of water at room temperature, though some inconsistences are present including high self-diffusion and some disagreement of the water structure in comparison with experimental data. Generally, TIP3P is used in a rigid variant, using for example the SHAKE algorithm to constrain the bond lengths, and H-O-H angle (constrained by including a fictitious H-H bond). It allows increasing the time-step during MD simulations to a typical value of 2 fs, needed to perform longer simulations. By contrast to the other FFs, in the CHARMM FF the TIP3P water model also plays a role during FF development, since interactions with TIP3P waters (the TIP3P geometry is also used for QM calculations) are used to derive point charges.

ADDITIVE CHARMM FF

CHARMM FF for proteins: history and the current state

The first additive CHARMM FF (version C4) was released in 1983, and similar to AMBER and GROMACS, was intended for proteins, while OPLS FF was initially created for organic liquids.[82] C4 was an united-atom model with an option to choose whether only aliphatic or all hydrogens are combined to heavy atoms. The empirical expression used for potential energy (Eq 2) contained the terms $E_b, E_\theta, E_\phi, E_\omega, E_{el}, E_{vdW}$, and E_{hb} for bonds, angles, dihedrals, improper torsions,

electrostatic interactions, van der Waals interactions, and hydrogen bonds, respectively:[82]

$$E = E_b + E_\theta + E_\phi + E_\omega + E_{el} + E_{vdW} + E_{hb} \text{ (Eq 1.2)},$$

where bonded terms are given by:

$$E_b = \sum k_b(r - r_0)^2 \text{ (Eq 1.3)}$$

$$E_\theta = \sum k_\theta(\theta - \theta_0)^2 \text{ (Eq 1.4)}$$

$$E_\phi = \sum |k_\phi| - k_\phi \cos(n\phi), \text{ where } n = 1,2,3,4,6 \text{ (Eq 1.5)}$$

$$E_\omega = \sum k_\omega(\omega - \omega_0)^2, \text{ (Eq 1.6)}$$

b_0, θ_0, ω_0 are the bond, angle and improper dihedral angle equilibrium values, respectively; and the k_x are the force constants. Force constants for bond, angle and stiff dihedral terms were obtained by fitting to vibrational data from experiments or from QM Normal Mode Analysis (NMA). Equilibrium values are typically adapted from crystallographic data,[82] with no or very small adjustment.

In C4 electrostatic interactions are modelled by Coulomb's law:

$$E_{el} = \sum \left(\frac{q_i q_j}{4\pi\epsilon(r)\epsilon_0 r_{ij}} \right) \text{ (Eq 1.7)}$$

where q, ϵ_0, r_{ij} are the atomic charge, the electric constant and the distance between atoms i and j, respectively; $\epsilon(r)$ is the distant dependent dielectric constant.

The vdW interactions were modelled by the Lenard-Jones (LJ) potential:

$$E_{vdW} = \sum \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \text{ (Eq 1.8)}$$

contain the repulsive (A_{ij}) and attractive (B_{ij}) parameters for the LJ 6-12 potential. Hydrogen bond term was present explicitly in the CHARMM FF only in the first version. Next versions of CHARMM FF did not explicitly include hydrogen bond terms, treating it implicitly by reducing the repulsive contribution of E_{vdW} term for the atoms involved in a hydrogen bond.

As a major step in the CHARMM development, CHARMM19 (version C19) was released in 1985 and was the last CHARMM FF employing a united-atom paradigm.[83] The C19 force field was popular until the end of the last century when all-atom version of the CHARMM FF became available. In C19, the dihedral torsion term was modified by adding a phase dependence with two phase values possible 0° or 180° allowing the force constant to be always positive (needed for implementation reasons

only). Polar hydrogens were treated explicitly in C19 and only hydrogens bonded to sulfur and carbon were still included to extended atoms (carbon atoms with larger vdW radii). In this update, the hydrogen bonds term was dropped, and a major revision of partial charges was accomplished to obtain hydrogen bonding interactions consistent with QM calculations of formamide–water interactions. The TIP3P water model was used and a scaling factor of 0.4 was applied to Coulomb interactions of 1-4 atoms and distance-depended dielectric parameter was still applied.[70,83] The modified TIP3P water model was retained for the following versions of the CHARMM FF, while AMBER and OPLS use the initial TIP3P model (without vdW sites on hydrogens).

With C19, the parameters enter the vdW interaction term in a different form:

$$E_{vdW} = \sum_{i>j} \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \text{Eq (1.9)}$$

in which ε_{ij} is the well depth and is the σ_{ij} radius at which the Lenard-Jones potential has a minimum (in contrast to A and B in the previous version). In addition, the LJ parameters for pairs of atoms i and j are constructed using:

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \; and \; \sigma_{ij} = \left(\sigma_i + \sigma_j \right) \text{Eq (1.10)}$$

As computer hardware was improving, CHARMM22[70] (C22) was released in 1991 and is considered to be the first all-atom protein FF and was used as a basis for all the following all-atom additive CHARMM FFs. The later attempts, except the introduction of the CMAP term, mostly focused on expanding parameters to other molecules. Thus, the CHARMM FF was supplemented with parameters for nucleic acids[84] and lipids.[85] Starting with this version (>22), the non-bonded terms were modified: the LJ atomic radius for atom pairs is also computed with the Lorentz-Berthelot combination rule, but as an arithmetic mean of the individual atomic radii. Obviously, it only reduces the values of vdW radii by two. However, the vdW parameters need to be redefined for the parent carbon atom, since all hydrogens are present in C22. For electrostatic term, the distance-depended dielectric parameter was replaced by a dielectric constant of unity. For the bonded part, the Urey-Bradley (UB) term (currently is considered as obsolete) was added to improve the agreement with vibrational spectra when a harmonic term alone would not adequately fit:

$$E_{UB} = k_{UB}\left(r_{1-3} - r_{1-3_0}\right)^2 \; Eq(1.11)$$

where k_{UB} is the force constant and r_{1-3_0} is Urey-Bradley equilibrium value, r_{1-3} is the distance between 1-3 atoms forming a valence angle. The parameterization of internal terms was based targeting experimental gas-phase geometries, vibrational spectra, and torsional energy surfaces completed with QM data. The bonded parameters for the peptide backbone were also optimized by fitting QM data for N-methylacetamide and the alanine dipeptide.[70] The C22 parametrization procedure extensively targeted experimental data in the condensed phase. In particular, the optimization included dipole moments, experimental heats and free energies of vaporization, solvation and sublimation, molecular volumes, and crystal pressures.[70] C22 fitting of partial charges relied on reproducing QM interactions energies and geometries between small fragments and a probe water molecule in vacuum testing different interaction sites. In addition to probe water interactions (modelled as TIP3P), the QM dipole moment was also included for neutral fragments only.

Another important methodological update was the addition of "CMAP"[86] (correction map) term for dihedral cross-term corrections in order to improve the accuracy of protein backbone conformational behavior during simulations. CMAP represents an error function between QM and MM potential energy surfaces, and mathematically is a 2-dimensional (2D) grid based correction (using cubic splines) for the backbone ϕ, ψ torsion angles, which allows to capture better the features of the 2D distribution of ϕ, ψ angles observed in structures deposited in the Protein Data Bank.[86,87] OPLS and AMBER FFs use the CMAP term to improve the conformational behavior of the protein backbone.[5,88]

The following updates of the FF, C27,[89,90] concerned parameters for nucleic acids and lipids and for carbohydrates (C35).[91] The results on simulations in longer-time scales indicated flaws of the FF. The identified shortcomings were remedied in the subsequent version, C36, released in 2012[92]. In parallel to backbone improvement (CMAP), C36 includes a systematic reparameterization of sidechains dihedrals (χ_1, χ_2). Sidechain dihedral terms were optimized by fitting to backbone-dependent QM PES scans, compared to crystal data from the Protein Data Bank[93] and followed by additional empirical optimization targeting NMR couplings for unfolded proteins.[69] The revised C36 parameters represent an improved FF for the treatment

of conformational sampling of the backbone with more accurate secondary structure propensities and also for the better sidechain sampling of χ rotamers.

As a significant interest appeared for intrinsically disordered, or partially ordered proteins that may be involved in cellular functions, a refined version of C36 was released in 2017[71], called C36m, with the CMAP term improved and also with atom pair-specific Lenard-Jones (L-J) parameters introduced (e.g. NBFIX in CHARMM nomenclature). The latter was needed to correct over-stabilized salt bridges in C36. The C36m is the latest version of the CHARMM FF (on the day of writing this memoire).

CHARMM FF for Small Molecules

Historically, classical force field models (CHARMM, Amber, and OPLS as example) were first created for important biopolymers like proteins, nucleic acids and lipids, presented above. However, for function these macromolecules need assistance from small molecules. For example, chemical reactions catalysed by enzymes frequently employ cofactors. The coverage of the much wider chemical space is needed to model these molecules, however in contrast to biological polymers (proteins, nucleic acids and lipids), which are assembled from a relatively small number of possible chemical blocks, small molecules have practically an infinite number of possible chemical compositions, making it challenging for a force field model. A significant effort was made to extend force fields to small molecules during the last decade, and as a result of this effort, force fields for small molecules were developed including the General Amber Force Field (GAFF)[94], CHARMM force fields for small molecules (CGenFF)[95], OPENFF[96], and OPLS-AA[97].

CGenFF force field was created to be used along with the standard CHARMM force field for proteins and nucleic acids and the CHARMM TIP3P water model. CGenFF explicitly aims at simulating small molecules in a biological environment described by the CHARMM classical force field. To this end, the same form of the potential energy function was adopted for CGenFF and mostly the same protocol for parameter optimization as for the standard CHARMM force fields. Differences with the standard force field for proteins include the absence of CMAP term, which is exclusively used to correct conformational dynamics around ϕ and ψ in the current C36m version of the CHARMM FF (*vide supra*); the lack of UB terms for newly developed parameters; the presence of explicit lone-pairs (LP) for halogen atoms.[95] Initially for CGenFF, two classes of model compounds were considered for the

parameter development. The first group includes a wide range of heterocycles, as these act as scaffolds of many important pharmaceutical agents; for the second group frequent and simple functional groups were considered that could be added to those heterocycles.

The development of force fields for small molecules, including CGenFF, is accompanied in parallel by advances in algorithms for automatically identifying atom types and generating (assigning) parameters for molecules from parameters explicitly optimized in other molecules. For example, for GAFF the Antechamber toolkit was created to allow the user to generate an Amber force field model for an arbitrary input molecule.[98] As for CGenFF, the *CGenFF* standalone program also available as web server (https://cgenff.umaryland.edu/) was designed to generate CHARMM topologies and parameters based on the CGenFF force field from atom connectivity.[99,100] Starting from the connectivity graph, which can be defined simply by using the distance matrix, the CGenFF program assigns atom types to atoms. The starting point is to discriminate atoms forming rings and acyclic structures. At the present, structures are considered as rings with less than 8 atoms; larger cyclic structures are treated as acyclic chemical moieties as they possess small ring strains.[99] Bond order is given to the CGenFF program in the mol2 file generated by OpenBabel[101] (or any other tool); the resonance structure, important to define partial charges, is resolved through calculations of the empirical score, which is a combination of the molecule net charge, atomic charges, potential (5,6,7-membered rings) and the number of aromatic rings. The atom types are defined through a set of condition rules, such as the number of protons bound to a carbon or nitrogen atom, valence and in-ring or acyclic atoms. Once atom types are assigned, the CGenFF tool defines parameters. These parameters may include already existing parameters, i.e. optimized previously for other molecules and included to the CGenFF, and new unknown parameters, assigned by analogy to the existing parameters. To this end, the score is computed between the missing term and existing terms as a sum of the scores for substituting the atom types defining parameters with the atom type in the missing parameter. The existing parameter with the lowest (best) score is then used for the missing one.[99]

The CGenFF program, in addition to parameters, also reports the associated fitness for parameter approximation, called penalty (P). The high penalty indicate

that the parameter may need further optimization, however these penalties only approximate the fitness by the analogy of the generated parameters with available parameters in CGenFF. Thus, in many cases explicit parametrization and validation are needed even for low-score terms. The parameters generated by the CGenFF program may serve as a good starting guess, facilitating initial stages of the FF development. The developed and validated FF model for novel molecules may be included to the standard CGenFF force field, which can be used for other similar groups and molecules in future.

The rest of this thesis is organized as follows:

The details of computational approaches used in this work are described in **Chapter 2**. The CHARMM FF functional form as well as the optimization strategy are discussed in detail. Methods for molecular dynamics simulations applied to test different aspects of the developed FFs are also described.

Chapter 3 presents the results of the FF development for nonstandard amino acids, which can be both, abundant in nature where they play a key role in various cellular processes and can be produced in laboratories. In this work, we have also extended the additive all-atom C36 to a large set of 333 nonstandard amino acids and CGenFF to 188 small molecules.

In **Chapter 4** the development of a force field model for the phenylalanyl group covalently linked to tRNA is described. This model is used to predict the structure of novel enzyme bound to tRNA in collaboration with the experimental group.

To obtain better FF bonded parameters we developed a new method based on the potential energy surface (PES) scans, presented in **Chapter 5**. This work was motivated by the fact that the conventional FF development methods can lead to the FF model significantly softer that the corresponding QM model under structural inconsistency, frequently present due to parameters transferred from existing molecules in the force field.

Typical drug design projects deal with large collections of small molecules, which are tested against a known receptor. *In silico* tests can be performed using a FF model to evaluate ligand-receptor interactions, requiring that such FF model being available. In this work, we are parametrizing a large collection of small molecules from the ZINC20 library, on order of hundreds of thousands, which can serve as fragments in drug design projects. This study is presented in **Chapter 6.**

This thesis is in part based on published papers: Chapter 3 and 4 are largely adapted from refs [102] and [103], respectively. The paper corresponding to Chapter 5 was submitted for publication, and Chapter 6 presents unpublished results. Finally, conclusions and perspectives for this work are given in **Chapter 7**.

Chapter 2
METHODS

CHARMM PARAMETRIZATION STRATEGY

Overall, the parametrization protocol adapted in this work follows the same steps of the C36 FF/CGenFF to ensure compatibility with the already available parameters, but with improvements to increase parameter transferability and stability (see **Chapter 5**). The potential energy (U) function from C36/CGenFF forcefields[104] was adapted without any modification:

$$\begin{aligned} U = & \sum_{el} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{vdW} \varepsilon_{ij} \left(\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right) + \sum_{bonds} K_b (b - b_0)^2 \\ & + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{UB} K_{UB} \left(r_{1-3} - r_{1-3;0} \right)^2 \\ & + \sum_{dihedral} \sum_{n=1}^{N} K_n \left(1 + \cos(n\varphi - \delta_n) \right) \\ & + \sum_{improper} K_\varphi (\varphi - \varphi_0)^2 + CMAP \end{aligned} \qquad \text{(Eq 2.1)}$$

The individual terms were described in the **Chapter 1**, while here the accent is made on the parameters. The electrostatic term described by Coulomb's law with atomic partial charges $\{q_i\}$ as empirical parameters; r_{ij} is the distance between atoms i and j. The partial charges are typically placed at the centre of nuclei, however, in special cases, the point charges can be placed at off-atomic-centre positions, representing for example lone-pairs. In the latter case, which is currently only applied in CGenFF, the out-of-atomic-centre is constructed using geometrical means and positions of parent atoms as a molecular frame. The geometrical parameters (the distance from the parent atom, for example) to find the place for LPs can be considered as empirical.

The vdW term is treated by the 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. The intramolecular or bonded part of the potential energy function is contributed by terms for the bonds,

valence angles, dihedral angles, improper dihedral angles, and selected Urey–Bradley terms, where b0, 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. In contrast to other bonded terms, a dihedral term, taking into account its importance for molecular conformations and dynamics, is represented as a Fourier series with N number of multiplicities to allow modelling complicated potential energy surfaces due to rotation around the dihedral angle. The current C36 FF uses multiplicities n = 1, 2, 3, 4 and 6 for a dihedral term, while multiplicity of n=5 is not generally recommended. The phase can take only two values δ_n = 0° or 180°, to reduce the total number of parameters for fitting. Note, that, in principle, with this convention the phase is not needed if the force constant can take also negative values. The CMAP term is a grid-based dihedral correction map (an error function) applied to the protein backbone, taken the importance of its conformational properties for the protein structure and dynamics, and which was not considered explicitly for optimization in this work (as described in **Chapter 3**).[87]

The optimization protocol is shown in Figure 1 of Chapter 3. In the first step, a geometry was produced for each molecule, if no experimental structure is available, to generate atom types and initial predictions for the parameters using the CGenFF program.[105] All parameters needed for the MM model for the molecule were separated into two groups: parameters that do not need adjustment, as indicated by the good analogy mitigated in the low penalty, and parameters that need further any adjustments, "missing" from the existing force field. In all cases, all partial charges were optimized with a few exceptions such as aliphatic and aromatic hydrogens (see below). Based on the "missing" parameters that need to be further optimized, the compound is further broken into simple model compounds that are "representative" for optimization. The order of the parametrization, i.e. steps of optimization of intramolecular and intermolecular terms, plays an important role and can lead to different sets of parameters and affect the quality of the force field model. Optimization starts with adjusting the intermolecular terms followed then by the optimization of intramolecular terms, since intramolecular interactions are also affected by intermolecular interactions. For example, the rotational energy of a hydroxyl group depends on the corresponding charges on the -OH group as well as the point charges on the parent molecule. The reverse dependence also exists, so typically the partial charges are tested and/or re-optimized with empirical CHARMM

structures after intramolecular terms were fitted, which implies an iteration process. The FF parameters are iteratively modified until reaching the specified convergence between QM and CHARMM calculations.

Prior to processing compounds with the CGenFF tool, hydrogens were added with OpenBabel[101]. The CGenFF program (should be distinguished from the CGenFF force field) is proprietary, some features of which can be accessed via https://cgenff.umaryland.edu (in our case, the access to the program was provided by Alex MacKerell). This tool performs atom typing and the assignment of parameters by analogy to the CGenFF force field in a fully automated fashion when provided a mol2 file.[99,100] PDB files containing the structures of molecules were converted to mol2 format with the help of OpenBabel[101] program. Along with coordinates, the mol2 file contains the complete information on the covalent structure including the bond order, which can be used by the CGenFF program. For the special case of nonstandard amino acids, which were described by a mixture of CGenFF and C36 atom types, atom types were manually corrected in CGenFF generated files.

When assigning parameters, the CGenFF programs also provides the "penalty" value, a crude estimation on analogy. A high value (P>10) for the penalty indicates that the given parameter has poor analogy with the molecules for which the FF is available. We subjected all parameters with penalties P>10 to optimization, while the low penalty parameters were not optimized. However, if the optimization of the high penalty parameters fails in specific molecular contexts, low penalty parameters were also optimized, as the penalty is only a rough estimation of analogy. If a bonded term was present in multiple molecules, it was optimized only in one model compound, and further then adopted to the other molecules. However, tests were performed in multiple molecules sharing substructures with the model compound in special cases. Typically, a model compound, in which the term was optimized, was chosen to have fewer atoms and preferably neutral among molecules sharing the term.

Prior to optimization of intermolecular and intramolecular parameters, the molecular geometries were fully optimized with QM to default tight tolerances. All QM calculations in this work were performed with Gaussian09[106]. Optimization was performed in vacuum using the MP2/6-31G* model chemistry for neutral and cationic molecules. For anionic molecules, the MP2/6-311G(d) model chemistry was applied. Since optimization was 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 during optimization to prevent the formation of the neutral form, dominant in vacuum, by the self-proton transfer to the carboxylic acid. In special situations where water probing of some atoms was not possible due to steric inaccessibility, a second geometry minimized with constraints of the molecule was used.

The parameters are first determined for small fragments, called model compounds, and subsequently transferred or adapted for the larger complete molecule. Large molecules are fragmented as to minimize the effects of "cutting" bonds and adding additional hydrogens, so that parameters optimized in smaller compounds can be used for larger molecule without any further adjustments. For this reason, fragmenting of molecules is performed by cleaving sites chosen between two acyclic saturated carbons to minimize effects on parameters due to cutting, i.e. by the smallest-possible perturbation of the system. The chemical structure of the model compound is completed by adding a proton to the acyclic saturated carbon on the cleavage site. The fragmentation of molecules and addition of missing hydrogens was accomplished manually using OpenBabel[101] or PyMOL software.[107]

INTERMOLECULAR TERM PARAMETRIZATION

The intermolecular parameters are optimized in accord with the standard CHARMM parametrization strategy. Optimization of LJ parameters is done in an iterative manner using expensive MD simulations to calculate condensed-phase properties (including densities and enthalpies of vaporization) which are compared with target experimental values. These iterations are done until a satisfactory agreement is reached. Since the results of these MD simulations also depend on partial charges, this protocol implies self-consistent iterations, where charges and LJ parameters are adjusted iteratively. Taking into account the importance of vdW parameters that affect all other parameters, creating new vdW types in the CHARMM FF is done by the laboratory managing the CHARMM FF (Prof. Alex MacKerell at Univ. of Maryland). In this work, no attempt to introduce new atom types or optimization of LJ parameters was made.

After vdW atom types and associated LJ parameters are defined for all atoms, partial charges are derived. Charges of aliphatic protons were not optimized and were set to +0.09*e* in accord with the standard CHARMM protocol. However, they were still probed by water interactions to allow a better charge density distribution on parent atoms. Aromatic CH protons also had the standard value of +0.115*e*. Charges of symmetrical atoms were set to have an identical value during the charge optimization. The charge of a group of atoms was also constrained to a net value. The QM target data included: i) interactions between the model compound and individual water molecules, including interaction energies and geometries; ii) the electrostatic potential (ESP); and iii) the dipole moment for neutral molecules.

QM water interaction energies and geometries were obtained as follows. For a compound, for which charge optimization is performed, the QM-optimized geometry at the previous step (section 2.1) is used. For the water molecule, the TIP3P model geometry is employed. Atoms of the molecule that can form hydrogen bonds were probed by individual water molecules. Aliphatic hydrogens were probed with a single water orientation; aromatic hydrogens were probed with two water orientations. Hydrogen atoms that can form hydrogen bonds were probed by at least four water orientations. In cationic molecules, only positively charged atoms were probed and in anionic molecules, only negatively charged atoms were probed, as interactions with probe waters with ionized molecules are strongly dominated by the monopole of the molecule (net charge) in vacuum.

The interaction distance between a selected atom of the compound and the water molecule in an idealized linear orientation, i.e. where the interaction with a probe water is likely to be the strongest, was optimized. In the CHARMM standard protocol, optimization and energy calculations with probed water molecules are done at the Hartree-Fock (HF)/6-31G(d) level.[70,95] For simulations in the condensed phase, the target QM data computed in vacuum are empirically corrected, which is also needed to correct the deficiencies of the low-level QM used (HF). QM water interactions in vacuum were scaled by factor 1.16 only for neutral polar molecules to account for the physical behaviour in the bulk phase; the QM minimum interaction distance is corrected by subtracting 0.2 Å for all polar interactions involving neutral compounds. The use of HF of QM theory is essential to maintain compatibility with the rest of the CHARMM FF, although higher levels of theory may produce more accurate hydrogen bond geometries and energies. In the case of sulfur atoms, the model compound-water interactions were calculated at the MP2/6-31G* level

including the basis set superposition error (BSSE) correction of Boys and Bernardi[108] and without applying standard scaling and offset rules.

The dipole moment defined by the charge distribution was used to provide additional target data for the optimization of the atomic charges for neutral molecules. The QM dipole moment was calculated in vacuum at the MP2/6-31G* model chemistry using the QM-optimized conformation.[95] The QM dipole moment was increased by 30% according to the standard CHARMM protocol, similar to the TIP3P water model which has the empirical dipole of 2.35 D, which is 30% higher than the experimental gas-phase value of 1.86 D.[95] Both, the magnitude and direction of QM calculated dipole moment were targeted in charge fitting.[109] In particular, we use a simple term in the target function given by: $\left|\boldsymbol{D}_{QM} - \boldsymbol{D}_{MM}\right|$, where $\boldsymbol{D}_{QM}$ and $\boldsymbol{D}_{MM}$ are the QM and CHARMM dipole moments, respectively.

Electrostatic potential (ESP) was also included to the QM target data. ESP was computed at the MP2/6-31G* model chemistry (MP2/6-311G(d) for anions) in vacuum, similar to the RESP protocol adopted for the Amber force field.[110] Partial charges in large molecules optimized only targeting water interactions may lack the robustness, i.e. very different charge sets can reproduce equally well interaction energies, since only water interactions with few hydrogen-bond donors and acceptors at the molecular surface are probed. This is especially critical for charged molecules, where the QM dipole moment is not included to fitting and only fewer probe water positions are considered due to the dominant effect of the net charge.

The charge optimization was accomplished with a C++ program based on Powell minimization algorithms from Numerical Recipes.[111] The partial charges were adjusted iteratively to reproduce the QM target data. For charge optimization, the fitness or target function was constructed and included the following terms: the Root Mean Square (RMS) deviation of minimum interaction energies and distances for water interactions; the magnitude of the difference vector between the QM and MM dipole moments for neutral molecules; the RMS deviation between the QM and MM ESPs; and a term restraining charges to the initial charges. The latter term was introduced to prevent large deviations from the starting guess for charges, similar to the RESP method. These terms are appropriately weighted in the fitness function with the weights given in Table 2.1.

The reproduction of water-compound interaction energies and geometries has the largest weight in the cost function of all terms. The different scaling applied to the optimization of partial charges are also summarized in Table 2.1. The contribution due to water interaction distances to the fitness function is the smallest, as it is mostly a function of LJ parameters rather than atomic charges.

Table 2.1. Weights in the fitness function for optimization of partial charges and scaling factors for QM data.

QM target data		Scaling of QM data	Target function term	Weight for the term in target function
Water interaction	Energy	$\times 1.16^{*}$	RMS between MM and QM	$10.0\ \mathrm{kcal}^{-1} \times \mathrm{mol}$
	Distance	$-0.2\ \text{Å}^{*}$		1.0 Å-1
Dipole moment		$\times 1.3^{*}$	Absolute difference between MM and QM	$3.0\ \mathrm{Debye}^{-1}$
Electrostatic potential (ESP)		no scaling	RMS between MM and QM	$1.0\ \mathrm{kcal}^{-1} \cdot \mathrm{mol} \cdot \text{Å}$

∗ scaling of the QM data and shifting of interaction distances were performed only for neutral molecules

BONDED TERM PARAMETRIZATION

Intramolecular interactions are modelled in the CHARMM FF by bond, valence angle, Urey-Bradley cross interaction, dihedral angle, improper dihedral, and CMAP term. In this work, the CMAP term was not considered for optimization.

Flexible dihedrals, also known as soft or rotatable, have a shallow potential energy surface having several minima that can be accessible during MD simulations at room temperature. To parametrize dihedral terms, PES scans were performed for each torsion, by adiabatically relaxing all other degrees of freedom, with the scanned dihedral angle constrained. In most of cases, a molecule has N rotatable dihedrals, and in this case, during the adiabatic scans all other rotatable dihedrals were also constrained to values in the minimum-energy structure. Note that this corresponds to the assumption that transitions in PES along rotatable dihedral angles are orthogonal and these degrees of freedom are independent. However, this assumption may break for ring structures, and multi-dimensional dihedral scans may be needed,

as it was done to parametrize the standard CHARMM FF for the dihedral angles in ribose.

Rotatable dihedrals were scanned in the range from -180° to 180° in 10° increments to cover a complete rotation. QM calculations were performed at MP2/6-31G* model chemistry (MP2/6-311G(d) for anions). CHARMM conformations were obtained starting from the geometries extracted from the QM scan, and by minimizing them with the CHARMM program[112]. No correction was applied to QM PES energies, in accord with the standard CHARMM protocol. During CHARMM minimization, a harmonic restraint with a large force constant of $5 \cdot 10^4$ kcal·mol^{-1}·rad^{-2} was applied on the target torsion, while other rotatable dihedrals were restrained to the values corresponding to the minimum energy geometry. The dihedral parameters were adjusted until a satisfactory agreement was achieved between the QM and MM surfaces for the low-energy regions, which are defined as PES regions with energies lower than 10 kcal·mol^{-1} from the minimum energy.

PES associated with non-rotatable dihedrals are typically characterized by a single minimum and high energy for small deformations. This type of dihedral is considered rigid/stiff and followed the same optimization method as for bond, valence angle, Urey-Bradley and improper torsion terms. As a major deviation from the standard CHARMM protocol, the stiff terms were optimized by adiabatic PES scans in my work. PES scans were performed for 7 equally spaced distortions for each internal degree of freedom. A similar method was used in CGenFF parametrization to determine force constants when the assignment of contributions of the internal coordinates to the vibrations was ambiguous, however, for CGenFF three-point PES scans were performed with deformations constant for all scanned degrees of freedom (We can name it as a *constant-max-deformation* PES scan).[109,113] In contrast, in our work, to ensure that only regions of PES sampled during MD simulations are parametrized, the value of increments were adjusted as in the previous work.[114] We name this method a *constant-max-energy* PES scan. In this simple method, the initial distortion increments from QM minimum energy structure were 0.06 Å, 4.0°, 25.0°, and 25.0° for the bond length, valence angle, dihedral and improper angle, respectively. While performing PES scans, even for relatively small deformations, energy may become very high and unreachable in MD simulations. For this reason,

the above initial values for the distortions are then corrected using the following equation:

$$\Delta x' = \sqrt{2\ \Delta E_{max}/k}\ \text{[Eq. 2.2]}$$

$$\text{with } k = 2\frac{(E(\Delta x) - E_0)}{\Delta x^{\,2}}\ \text{[Eq 2.3]}$$

where 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, respectively. ΔE_{max} defines the highest energy on scanned PES. To optimize each bonded term, the seven equally spaced points used for PES were in the range of $x \in [x_0 - \Delta x',\ x_0 + \Delta x']$, including the minimum energy structure at $x = x_0$. In Equation 2.3, $\Delta E_{max} = 2.0\ \text{kcal} \cdot \text{mol}^{-1}$ was used. All PES scan energies are limited by this ΔE_{max} value, hence the name *constant-max-energy* PES scan. PES scans were performed at the same MP2/6-31G* model chemistry and MP2/6-311G(d) for anions with G09 software[106].

Similar to soft dihedral optimization, each conformation for the CHARMM 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, PES adiabatic scans were performed with CHARMM program using a new set of CHARMM parameters.

To adjust the different CHARMM bonded parameters simultaneously, the C++ program described for charge optimization was used. The program is based on Powell minimization algorithms from Numerical Recipes.[111] The target function for bonded terms optimization included RMS deviation between QM and empirical PES energies, RMS deviation between QM and CHARMM geometries; and restraints to the initial set of parameters provided by the Penalty term of CGenFF program predictions. In addition, the weighted RMS deviation between cartesian components of QM and CHARMM forces was added to the target function. At each optimization iteration, the empirical PES scans were performed with the CHARMM program[112] using a updated set of parameters. The bonded parameters were iteratively modified until the target function could not be reduced further.

MOLECULAR DYNAMICS (MD) SIMULATIONS

MD simulations of protein complexes were performed starting from crystal structures retrieved from the Protein Data Bank (PDB). To prepare the MD model, protonation states of residues were determined with the PROPKA[115,116] tool, that assigns pK_a's to titratable residues using an empirical scoring function. In the case of histidines, the protonation was assigned by visual inspection and ideal stereochemistry. Water molecules present in protein crystal structures were preserved in the MD model. 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. With this distance, voids, which can be hydrophobic in nature, inside the protein are not filled with waters from the overlaid water box. The size of the water box was chosen so that the sides of the box were at least 10-12 Å away from any atom of the protein. An appropriate number of potassium or chloride counterions was included to render the system electrically neutral.

The system preparation was done with the CHARMM[112] program of version c41b1, while MD simulations were performed with the NAMD package version 2.2[117], typically running on GPUs for efficiency.[117] Periodic boundary conditions were applied, and the entire box was replicated periodically in all directions. All long-range electrostatic interactions were computed efficiently by the particle mesh Ewald method[118] using a real-space cut-off of 11 Å. Long range electrostatic forces were evaluated every 4 steps, while short-range non-bonded interactions were computed at each step. All vdW interactions were truncated at the distance of 11 Å with a smooth switching function. MD simulations were performed at NPT ensemble at constant room temperature and pressure, after 200 ps of thermalization. Constant pressure was maintained using the Berendsen pressure bath coupling[119] with the relaxation of 500 fs, the compressibility parameter of liquid water, by rescaling coordinates of atoms. Constant temperature was maintained by coupling to a heat bath with a room temperature by correcting forces as implemented in the NAMD program.[117]

We used two types of the system setup depending on the protein size and goals of the study. In one setup, considered complete setup, all protein residues were present in the MD model. In this case, the centre of mass of the protein heavy atoms was weakly restrained to the initial position of the system by a harmonic potential

with a force constant of 0.1 kcal × mol^{-1} × $Å^{-2}$ to prevent the drift of the protein in MD simulations, so that the protein atoms interact with the same cell unit.

The second setup is suited when one is only interested in the structure/dynamics of "central" region.[120,121] The centre of interest can be a catalytic site or a nonstandard amino acid as in **Chapter 3**. The residues situated within 24 Å around the centre of interest were maintained in the MD model, while the residues beyond the 24 Å radius sphere were removed. To preserve the net charge of the system, which can be further neutralized by adding counter ions, the truncation was done based on the CHARMM "groups", which are groups of atoms in a residue having a net charge. Similar to the first system setup described above, the spherical protein region was overlaid by a water box with its edges at least 10-12 Å away from protein atoms. The system was neutralized with appropriate number of counterions. The rest of the setup follows the complete protein setup presented above, except restraints acting during MD simulations. For the spherical region, no restraint to the centre of mass was applied, but atoms between 20 and 24 Å from the sphere's centre were harmonically restrained to their experimentally determined positions.

In all simulations, the C36 force field was used for the protein[69,110] and the CHARMM standard model for water (modified TIP3P).[70,80,83] The nonstandard groups/molecules such as nonstandard amino acids were modelled using the force field parameters specifically developed in this work.

Chapter 3
ADDITIVE CHARMM FORCE FIELD FOR NONSTANDARD AMINO ACIDS

Protein activity, participating in almost every process within cell, whether is catalytic or not is defined by its 3-dimensional structure, which in it turn defined by the composition of amino acids. There are 20 amino acids that have their own designated codons and 2 additional amino acids, selenocysteine and pyrrolysine, for which special coding mechanisms exist depending on the concerned organism, also encoded by the translational machinery. Thus, there are 22 *standard* amino acids.

In contrast to the standard amino acids, there are many more nonstandard amino acids that are not encoded in the genetic code. However, they are naturally abundant as post-translational modifications (PTM), as intermediaries in metabolic pathways, and born due to oxidative stress reactions. Additionally, they can be of artificial origin and synthesized for specific applications.

In this work, we selected a set of 333 nonstandard amino acids, which did not have a model in the CHARMM FF, for force field development. It includes 198 non-canonical amino acids from the SwissSidechain database of amino acids[122] and an additional 135 most frequent nonstandard amino from the Protein Data Bank (PDB).[123] The latter were selected based on the survey of PDB. In particular, each selected amino acid was present in PDB structures of at least two significantly different proteins (with the sequence identity < 90%). At the moment of writing of this thesis, there are at least 16,000 structures in PDB containing amino acids from the selected set.

Nonstandard amino acids can be classified in two group, first, those which share standard a backbone group and, second, those that have backbone modified groups. The standard amino acids, except proline and glycine, have different sidechains and share a backbone constituted out of a carboxyl (-COOH) and an amine group ($-NH_2$) attached to carbon α (Cα). Considering the type of modification they have, in the selected nonstandard amino acids there are 42 that present modifications at the level of backbone and the remaining 291 amino acids have modified sidechains with a standard backbone. Two important residues were considered only as

standalone ligands as they are not present in polypeptides in the PDB. Amongst the amino acids with modified backbone, four of them presented the modification at the level of N-terminus and were optimized only as N-termini peptides.

Examples of a sidechain modified amino acid included in this work are 2-amino-3-cyclohexyl-propionic acid (ALC) and seleno-methionine (MSE). ALC was included in the peptide WW61 computationally designed as an inhibitor for amyloid fibril formation. Amyloid fibril formation is associated with pathologies as Alzheimer's disease.[124] MSE has antioxidant activities and play a role in the formation and recycling of glutathione.[125,126] Also, MSE can be incorporated into proteins, replacing methionines, and it is used to help resolve the structure of proteins by X-ray crystallography using single- or multi-wavelength anomalous diffraction.[127] Backbone modified amino acids considered for the FF parametrization include the example of the chromophore (GYS) of the green fluorescent protein. The green light emitting fluorophore is the result of cyclisation of serine, tyrosine and glycine. MDO, another backbone modified amino acid, is the result of serine alanine and glycine cyclization and ensures the catalysis of the first reaction in histidine catabolism by histidine ammonia-lyase enzyme. MDO was also used to study the GFP chromophore biosynthesis.[128]

For sidechain modified amino acids, dipeptides were created by adding an acetyl group at the N-terminus and a N-methylamide group at the C-terminus. For the modified backbone amino acids tripeptides were created with the sequence ALA-X-ALA, with X the nonstandard amino acid as shown in Figure 3.1. The first alanine is acetylated and the last one is methylated allowing the parametrization of the dihedrals of modified backbone. The peptides were created to parametrize the torsion term corresponding to the rotation around the bond $C_\alpha - C_\beta$ (χ_1) for sidechain modified amino acids and is similar to the C36 parametrization protocol. In C36, dihedral potentials were optimized against QM energies from dipeptides and NMR data from unfolded proteins. The same dipeptide model was used for Gly, Pro and Ala for sidechain dihedral optimization, considering that such dipeptides are representative of the amino acid sidechains in the local environment of peptide backbone.[69] Alanine was representative of the remaining standard amino acids, excluding proline and glycine. This is the reason why the sidechain modified amino acids were also parametrized in the dipeptide form. Similarly, the CMAP term optimization was performed by targeting QM data for short peptides of poly-alanines

for standard amino acids, except Pro ang Gly. We adapted the existing CMAP from alanine to all sidechain modified nonstandard amino acids.

For backbone modified amino acids, the CMAP correction term was not specifically developed, however with the accurate optimization of dihedral parameters of the backbone was performed using tripeptides. During optimization of the rotatable dihedral torsions the QM target constituted of 36-point potential energy surface scans similarly to χ_1 and χ_2 of sidechain modified amino acids.

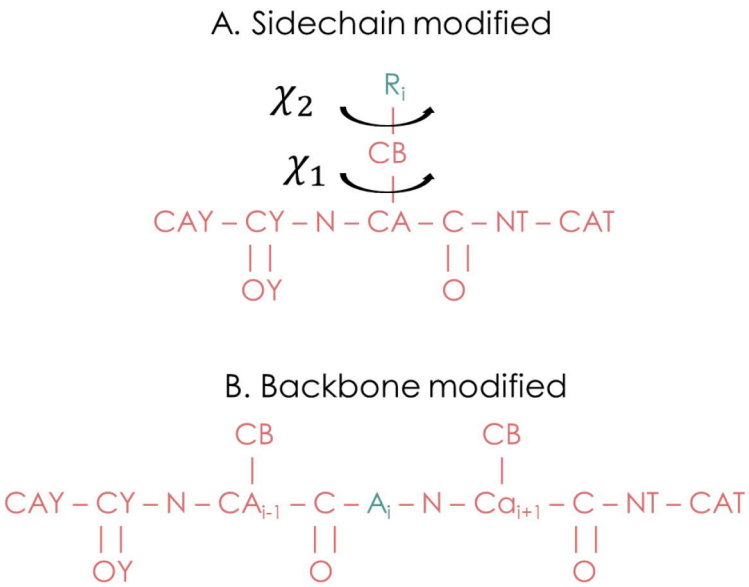


Figure 3.1. Model compounds to optimize bonded terms for (A) sidechain modified amino acids; the torsions χ_1 and χ_2 are shown by a black arrow; for (B) backbone modified amino acids were incorporated into tripeptides with the sequence ALA-X-ALA, with X the nonstandard amino acid.

For FF parametrization, the D- and L- stereoisomers were considered for 61 residues. The most important protonation and tautomeric states at physiological pH of 7 were also included. The pKa values and tautomeric forms were determined with MarvinSketch software version 19.19.[129] In total, 406 distinct forms were selected in which the atom names for heavy atoms were retained from Protein Data Bank, that were actually defined in the PDB Chemical Component Dictionary (CCD).[123,130] Hydrogen atom names were assigned according to the parent heavy atom to which they are bonded. Parameters were optimized using 188 small compounds representing the totality of functional groups present in the nonstandard amino acids.

The coordinates for peptides and small compounds were constructed in PyMOL software[107] from experimental PDB structures and the names of constructed

atoms had been taken from CHARMM FF. Once the PDB files together with the atom names for the molecules were created, I obtained atom types and initial guesses for the parameters. An important tool in this sense is the CGenFF program at https://cgenff.umaryland.edu/. It performs atom typing and assignment of parameters by analogy to CGenFF in a fully automated fashion when provided a mol2 file.[99,100] PDB files containing the structures of molecules were converted to mol2 format with the help of OpenBabel[101] program and CGenFF program was used for the assignment of atom type and initial parameters. For the nonstandard amino acids, I ensured that the backbone atom types and associated parameters are from the C36, while the sidechains up to the Cβ atom have CGenFF atom types if no modifications are present at the level of backbone. The amino acids with backbone groups different from the standard backbone have CGenFF atom types and parameters for all atoms as indicated in Figure 3.1.

When assigning parameters, the CGenFF programs also gives a "penalty" value. A high value for the penalty indicates that the given parameter has poor analogy with the molecules for which the FF is available. Further, we considered that all penalties higher than ten require optimization, while the low penalty parameters will be adopted as they are. However, if the optimization of the high penalty parameters fails in specific molecular contexts, low penalty parameters will also be optimized.

If a bonded term was present in multiple molecules, it was optimized in a single molecule and then adopted to the other molecules. The molecule, in which the term was optimized, was chosen to be the smallest in size and preferably neutral compared to the other molecules sharing the term.

After optimization of all the parameters typical CHARMM topology and parameter files were created. CHARMM optimized geometries of nonstandard amino acids were used to generate internal coordinates tables for the topology files. Additionally, we performed testing of parameters to ensure lack of parameter duplicates and also detect any missing parameters in these files. First, we generated pentapeptides, Ala-Ala-X-Ala-Ala, with CHARMM program directly from topology and parameter files for each nonstandard amino acid and then the structures were minimized with CHARMM. Tests also included glycine and proline as flanking residues, since their backbone has different CMAP from alanine. For the final test,

we generated a single chain containing all optimized amino acids separated by alanines, i.e. having the sequence: ...-X_i-A-X_{i+1}-....

To validate the created parameters, we performed MD simulations of proteins containing nonstandard amino acids. While the validation of the CHARMM FF for the standard amino acids was done using different experimental data, including spectroscopic and NMR, in this work, the validation step was mainly done using experimental structures available from PDB. We note that such comparison of modelled structures with experimental structures is not straightforward, as MD simulations are performed at different conditions as crystallographic experiments. For example, X-Ray experiments are frequently performed at cryo-temperatures, while modelling using the FF model is often done at temperatures close to the room temperature. In principle, modelling at cryo-temperatures could be done, however, due to low kinetic energy structural fluctuations are very limited. Other factors include: dewatering of crystal structures, additional ligands present, and crystallographic packing effects. Overall, such comparison, while being a good test on force fields, especially in the case of absence of other experimental data available, should be done with additional care.

The validation the force field for the nonstandard amino acids was performed by MD simulations of 20 different proteins containing either sidechain or backbone modified amino acids. During the simulations, the conformations of protein heavy atoms as well as conformations of the nonstandard amino acid fluctuated around the experimental structures in all systems. The average deviation of heavy atoms 10 Å around the nonstandard amino acid from the experimental was only 0.4 Å on average. Moreover, the torsion angles of important backbone and sidechain dihedrals of the nonstandard amino acids in the simulations have a low deviation from dihedral values in the crystal structures of only 5.4°. The hydrogen bonds with the nonstandard amino acid present in the experimental structure are also maintained in the simulations. The ensemble of given results demonstrates the accuracy for both, the optimized charges and bonded terms parameters, and that the forcefield developed for nonstandard amino acids performs as well as CHARMM force field for standard amino acids.

In the next part of this chapter the published work for nonstandard amino acids parametrization is included.[102] Additional information on structures and

nomenclature of nonstandard amino acids considered in this work can be found in Appendix for this thesis.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

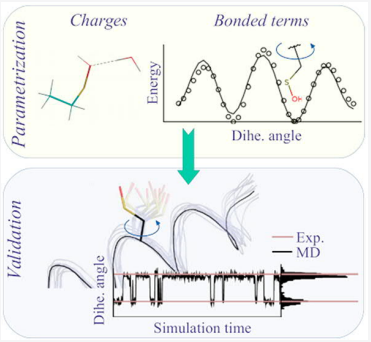
Additive CHARMM36 Force Field for Nonstandard Amino Acids

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

Cite This: J. Chem. Theory Comput. 2021, 17, 3554−3570 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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

Received: March 12, 2021
Published: May 19, 2021

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

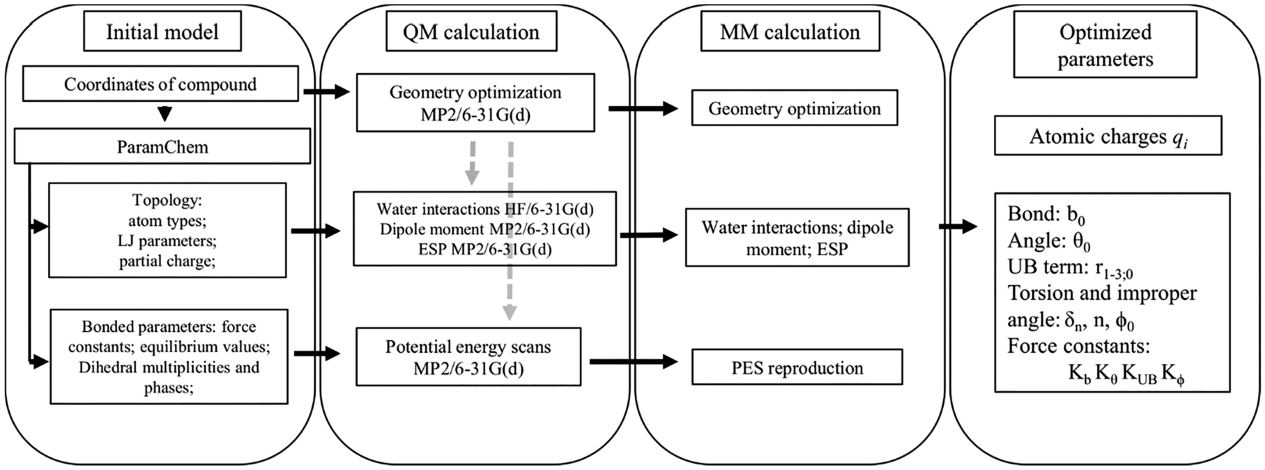


Figure 1. Workflow of force field parametrization. For anionic species, the MP2/6-311G(d) model chemistry was used for optimization and potential energy surface (PES) scans of the compounds.

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] Seventeen 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.

MATERIALS AND METHODS

CHARMM Potential Energy Function. The potential energy function of the nonpolarizable 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

$$U = U_{\text{inter}} + U_{\text{intra}} \tag{1}$$

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

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

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, and $R_{\min,ij}$ 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]

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

The intramolecular or bonded part of the potential energy function in eq 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

$$\begin{aligned} U_{\text{intra}} = {} & \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 \\ & + \sum_{\text{Urey−Bradley}} K_{UB}(r_{1-3} - r_{1-3;0})^2 + \sum_{\text{dihedral}} \sum_{n=1}^{N} K_n(1 + \cos(n\varphi - \delta_n)) \\ & + \sum_{\text{improper}} K_\varphi(\varphi - \varphi_0)^2 + \text{CMAP} \end{aligned} \tag{4}$$

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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

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: 0° or 180°. 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 (PDB) 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.

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 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 Lennard−Jones 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 (ESPs) 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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

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 45° or 90° of the water probe rotation for polar atoms and one or two orientations for nonpolar 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 was used 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 $kcal^{-1}$·mol·Å) was kept small relative to the weights for water-interaction (10.0 $kcal^{-1}$·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 nonrotatable. PES associated with nonrotatable 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 5×10^4 $kcal{\cdot}mol^{-1}{\cdot}radian^{-2}$ 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 $kcal{\cdot}mol^{-1}$ 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 nonrotatable dihedral angles were optimized using the following protocol. The initial guess

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

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

$$\Delta x' = \sqrt{2\Delta E_{max}/k} \tag{5}$$

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 $x = x_0$. In eq 5, 2.0 kcal·mol^{-1} was used for ΔE_{max}. 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 5×10^4 kcal·mol^{-1}·Å^{-2} or 5×10^4 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. The 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 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 four steps, while short-range nonbonded 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 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 ns.

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 the Supporting Information. This set of residues includes 198 amino acids from the SwissSide chain 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 the 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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

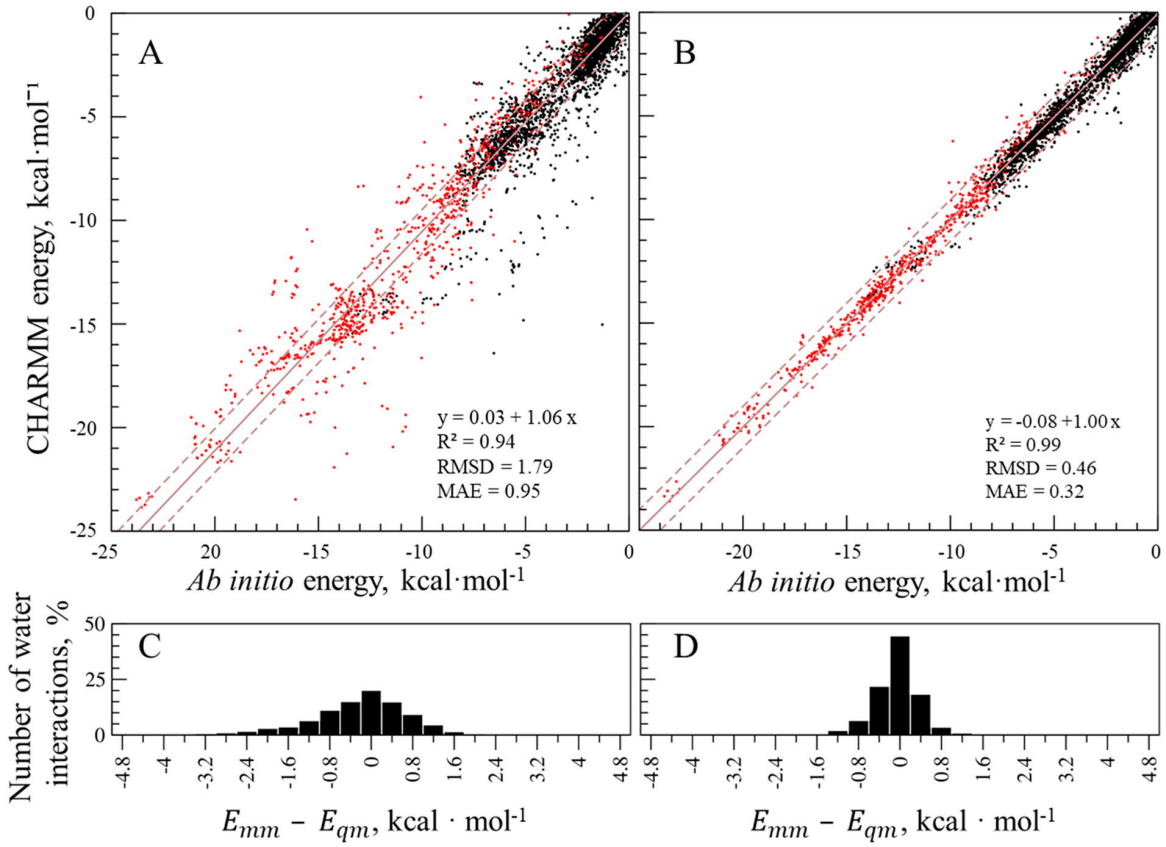


Figure 2. Corrected QM and CHARMM water interaction energies for the compound-water monohydrates. The CHARMM energies were computed using A) the initial ParamChem charges and B) the optimized atomic charges. C) and D) Percentage of water interactions vs energy deviation in A) and B), respectively. Interaction 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.

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 Materials and 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 the Materials and Methods section. This includes 906 probe water-model compound interactions for ionized compounds and 2951 for neutral compounds.

Figure 2 compares QM and MM interactions energies. The statistics for water-compound interactions for all compounds are given in Table 1. Empirical and *ab initio* interaction energies and distances are given in Tables S3−S678 in the Supporting Information. The RMS deviation for interaction energies with the initial ParamChem and optimized charges is

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 optimal/initial	MAE optimal/initial
norm of μ [a]	141	0.59/1.72	0.41/1.28
direction of μ [b]	141	2.4/16.3	5.1/32.7
water-solute E_{int} [c]	3857	0.46/1.79	0.32/0.95
water-solute d_{min} [d]	3857	0.20/0.66	0.16/0.28
φ_{elec} [e]	195	2.38/4.19	1.96/3.39

[a]The magnitude of the dipole moment (μ) is given in Debye. [b]Angle (deg) 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 $\sum\varphi_i p_i / \sum p_i$ 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]Probe water-model compound interaction energies are in kcal·mol^{-1}. [d]Probe water-model compound interaction distances are in Å. [e]Electrostatic potential is in kcal·mol^{-1}·Å^{-1}.

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

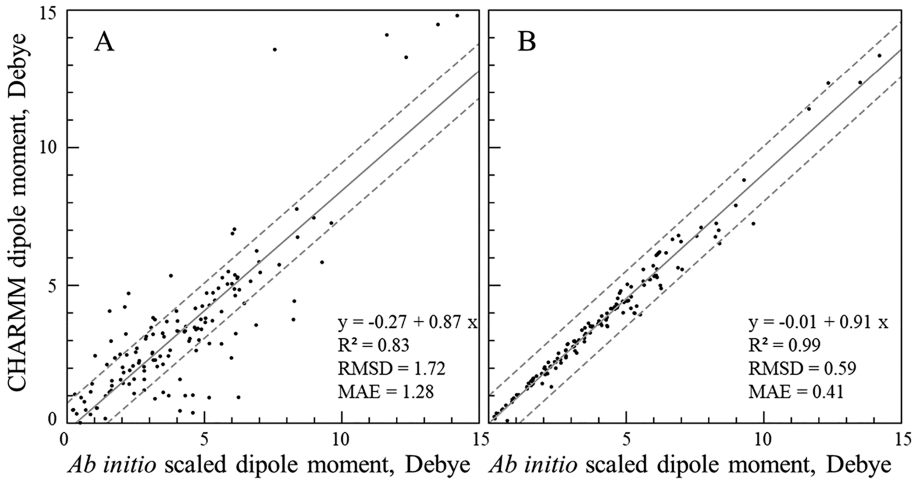


Figure 3. Comparison between the scaled QM (increased by 30%) and CHARMM dipole moments for 141 neutral model compounds. The CHARMM dipole moment was computed using A) the initial ParamChem charges and B) the optimized atomic charges. The linear regression line is shown by the solid line; the dashed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

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]

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 32.7° and 5.0° 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 8.5° 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 10° and larger than 10° 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 16.3° and 2.4° with the initial and optimal charges.

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 potentials 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]-

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

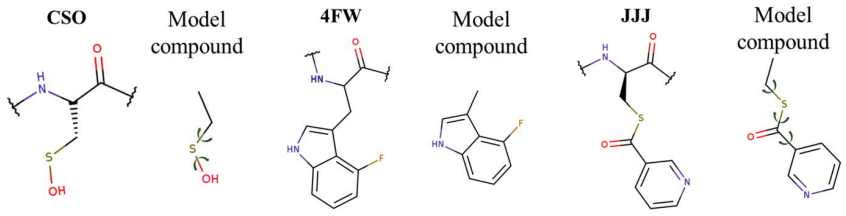


Figure 4. Model compounds used to parametrize 4-fluorotryptophane (4FW), *S*-hydroxycystein (CSO), and *S*-(pyridin-3-ylcarbonyl)-L-cysteine (JJJ). The rotatable dihedral angles parametrized in this work are indicated by arrows.

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 to the phosphorus 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 phosphorus 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 phosphorus 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 of 0.6 D is 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° and 7.7° 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.

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 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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

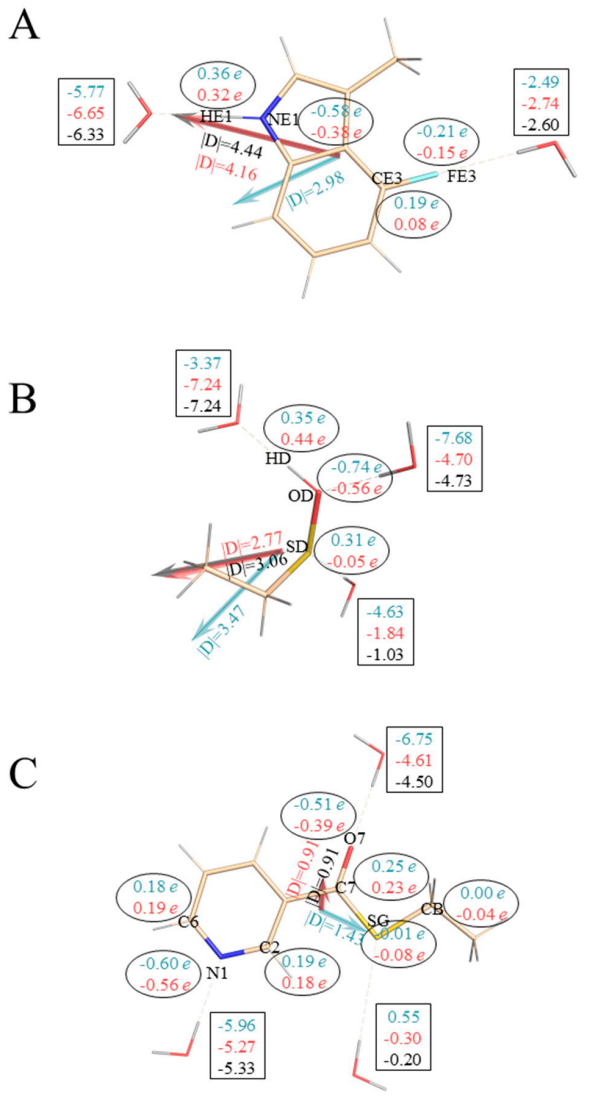


Figure 5. Selected water interactions with model compounds used to parametrize (A) 4FW, (B) CSO, and (C) JJJ. The water-compound interaction energies are given in the rectangular box: MM interaction energies are computed with the initial and optimal charges, and QM interaction energies are shown in blue,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

to 0.06 kcal·mol^{-1} after optimization from initial 0.63 kcal·mol^{-1}.

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 nonrotatable 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 180°. 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 1.6° and 1.4°, respectively. For torsions, the RMS deviation is 5.8° and 4.6° 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 optimal/initial	RMSD optimal/initial
RMSD (Å)[a]	189	0.14/0.18	0.18/0.24
bond (Å)	3519	0.015/0.016	0.020/0.023
angle (deg)	5968	1.4/1.6	1.9/2.7
dihedral (deg)	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 nonrotatable 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 and 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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

Table 3. Comparison between Empirical and *Ab Initio* Energies of PES Scans

term	N terms[a]	N points[b]	RMSD optimal /initial[c]	MAE optimal/ initial[d]	R optimal/ initial[e]
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]RMS deviation between QM and MM energies. [d]Mean absolute error. [e]Linear correlation, *R*.

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.

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 360° 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 180°. 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

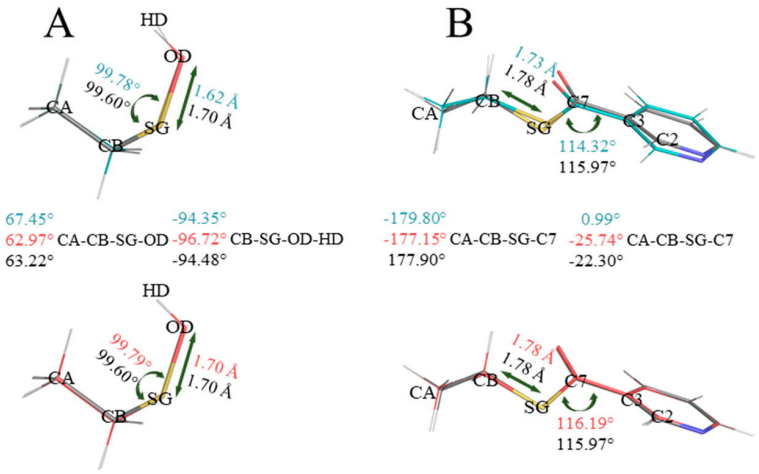


Figure 6. Comparison between QM and MM geometries of (A) CSO and (B) JJJ. The superposition of the QM structure and the structure optimized using the initial and optimal parameters is shown in the upper and bottom panels, respectively. The values for selected degrees of freedom are also given for the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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

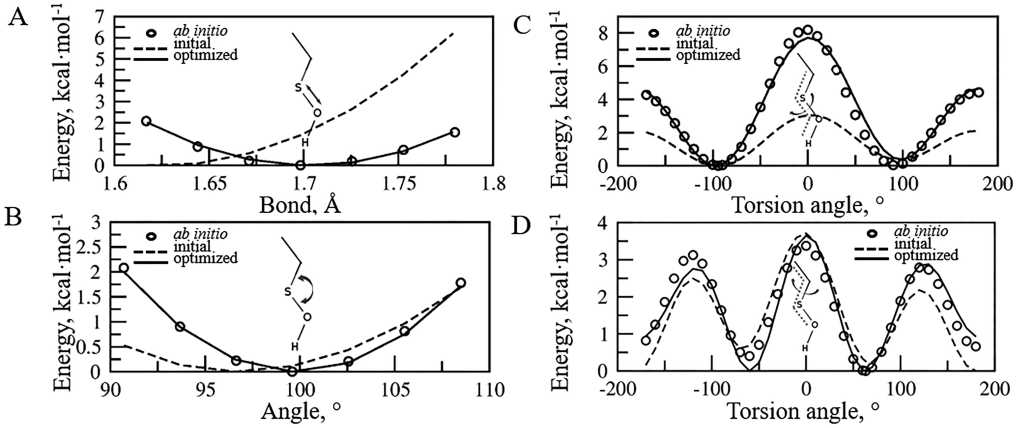


Figure 7. PES scans for selected degrees of freedom in the CSO model compound. The PES scan was performed for a bond (A), for a valence angle (B), and for the dihedral angles associated with the rotation of the hydroxyl group (C) and the sulfenic (D) group, respectively. The arrows and dotted lines indicate 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.

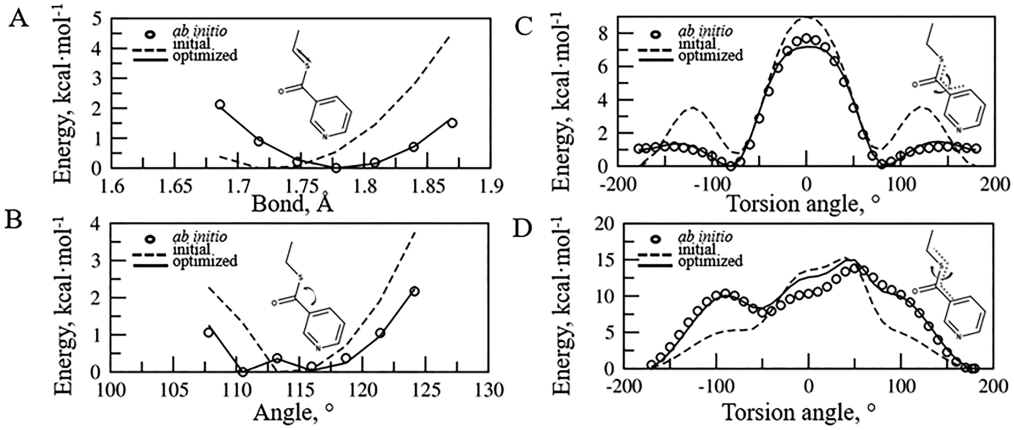


Figure 8. PES scans for selected degrees of freedom in the JJJ model compound. The PES scan was performed for a bond (A), for a valence angle (B), and for the two dihedral angles adjacent to the phenyl ring (C and D). The arrow and dotted lines indicate the corresponding degree of freedom along which 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.

can be seen in Figure 6, demonstrating that ParamChem provides a very good guess for these parameters.

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 Materials and 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 the 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 is 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 of the sulfenic group, the local minimum at −60° is ∼0.4 kcal·mol^{-1} higher in energy relative to the minimum at 60°, 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 0° or 180°, 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 nonharmonic 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.

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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

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

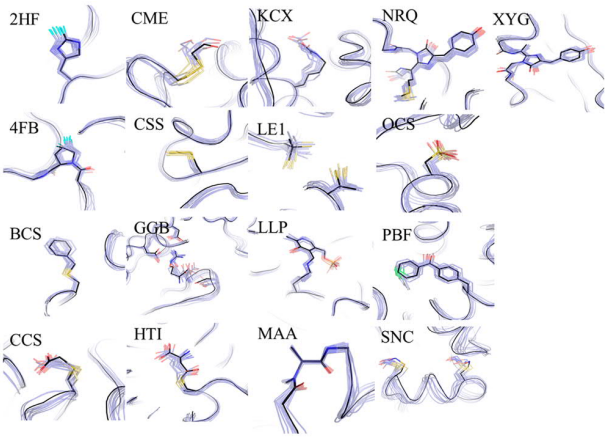


Figure 9. Comparison of structures from MD simulations (in gray) with the experimental structures (in color). Ten snapshots were taken every 10 ns from 100 ns MD simulations and superimposed on the experimental structure using the protein backbone atoms.

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 Tables 4 and S870. The RMS deviation

Table 4. Root-Mean-Square (RMSD) Deviation in Molecular Dynamics Simulations

amino acid	PDB ref	RMSD (Å) backbone[a]	RMSD (Å) backbone[b]	RMSD (Å) residue[c]
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.

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 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

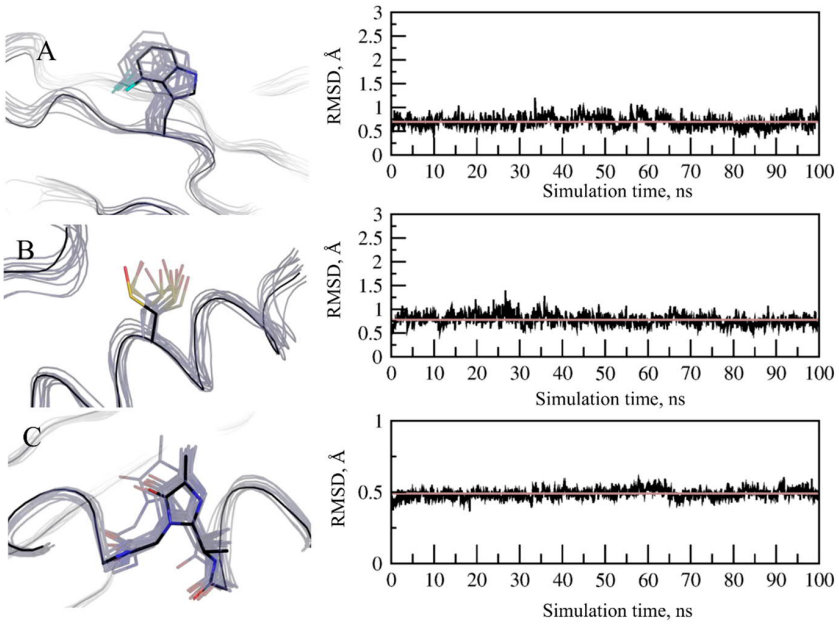


Figure 10. Comparison of structures from MD simulations (in gray) with the experimental structures (in black) for (A) 4FW, (B) CSO, and (C) MDO (PDB access codes: 6SZZ, 6Q00, and 1IYF, respectively). Ten snapshots were taken every 10 ns from 100 ns MD simulations and superimposed on the experimental 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.

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

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.

Tables 5 and S871 summarize selected nonbond 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 5.4° and the RMS deviation of 11.0°. The largest deviation of 55.5° 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 positions of these atoms were 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.

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 the Supporting Information. Details of the optimization of the parameters associated with CSO and 4FW were presented

Table 5. Selected Average Nonbond Distances (Å) in MD Simulations of Proteins with the Nonstandard Amino Acids

residue	atom pair[a]	X-ray str	MD simulation[b,c]	abs diff[c]
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 parentheses are the RMS fluctuations. [c]MD simulations were performed with the initial and optimized parameters, respectively.

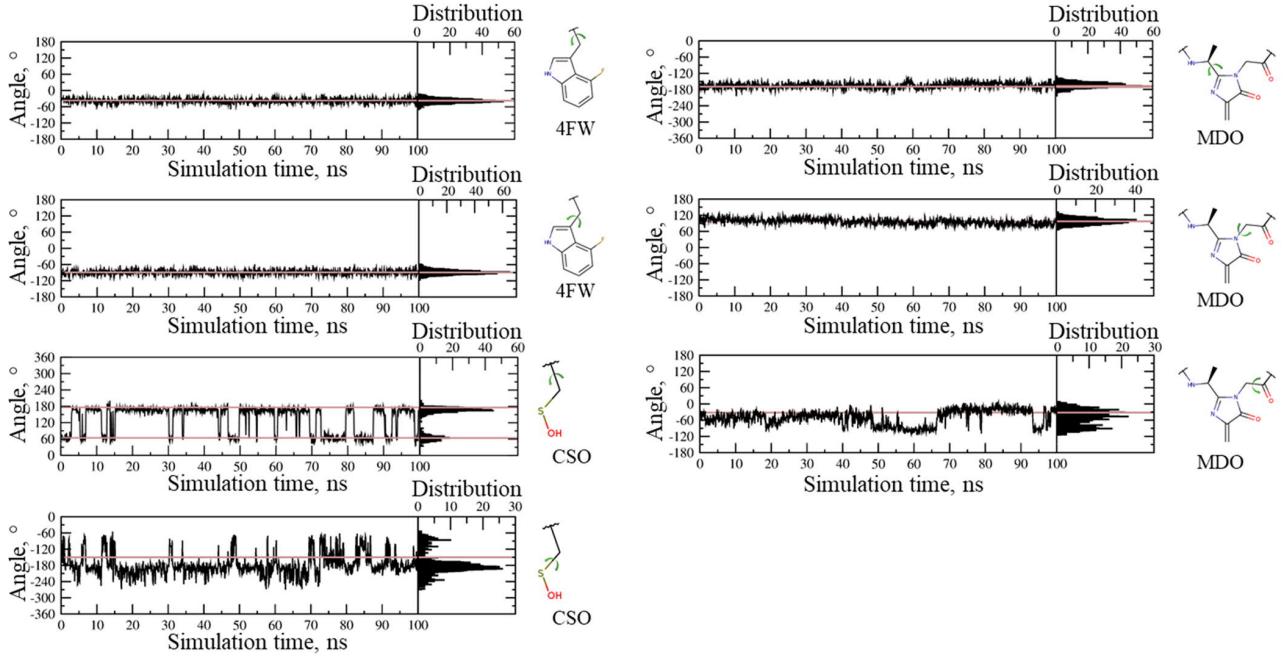


Figure 11. Rotatable dihedral angles in MD simulations of protein complexes with 4FW, CSO, and MDO (PDB access codes: 6SZZ, 6Q00 and 1IYF, respectively). The dihedral angle is shown by the arrow; the experimental value is shown by the solid gray line; the right panels show the distribution of the dihedral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

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554–3570

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.

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 are 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 63.6° and 164.3° (models 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 63.2° and 176.1°, and 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 (64.6° and 170.1°); however, the conformation with χ1 of ∼60° 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° by 1.2 kcal·mol^{-1} relative to 60° 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[b]

residue	dihedral	X-ray	MD[a]	abs diff[a]
MDO	CB−CA1−C1−N3	−168.7	−160.9 (10.7)/−164.4 (10.3)	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 (40.7)	36.6/28.4

[a]MD simulations were performed with the initial and optimized parameters, respectively. [b]RMS fluctuations are given in parentheses.

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 not only for the non-hydrogen atoms of the nonstandard 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 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 nonbond 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.

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

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 nonbonded 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 will be available from the MacKerell lab web page (https://mackerell.umaryland.edu/) and CHARMM-GUI (http://www.charmm-gui.org),[15,69] facilitating their utilization in a range of molecular simulation software packages.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00254.

> 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 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Alexey Aleksandrov − *Laboratoire d'Optique et Biosciences (CNRS UMR7645, INSERM U1182), Ecole Polytechnique, Institut Polytechnique de Paris, F-91128 Palaiseau, France;* orcid.org/0000-0002-8150-3931; Email: alexey.aleksandrov@polytechnique.edu

Alexander D. MacKerell, Jr. − *Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland 21201, United States;* orcid.org/0000-0001-8287-6804; Email: alex@outerbanks.umaryland.edu

Authors

Anastasia Croitoru − *Laboratoire d'Optique et Biosciences (CNRS UMR7645, INSERM U1182), Ecole Polytechnique, Institut Polytechnique de Paris, F-91128 Palaiseau, France*

Sang-Jun Park − *Departments of Biological Sciences, Chemistry, Bioengineering, and Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States;* orcid.org/0000-0002-7307-3724

Anmol Kumar − *Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland 21201, United States*

Jumin Lee − *Departments of Biological Sciences, Chemistry, Bioengineering, and Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States;* orcid.org/0000-0002-1008-0118

Wonpil Im − *Departments of Biological Sciences, Chemistry, Bioengineering, and Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States;* orcid.org/0000-0001-5642-6041

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00254

Notes

The authors declare the following competing financial interest(s): A.D.M. is co-founder and CSO of SilcsBio LLC.

■ ACKNOWLEDGMENTS

This work was supported by grants ANR-18-CE44-0002 to A.A., NIH R01GM138472 to W.I., and NIH R35GM131710 to A.D.M. This work was performed using HPC resources from GENCI-CINES (Grant 2018-A0040710436).

■ REFERENCES

(1) Böck, 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.–Online* **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, A. D., Jr CHARMM Additive and Polarizable Force Fields for Biophysics and Computer-Aided Drug Design. *Biochim. Biophys. Acta, Gen. Subj.* **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.* **2012**, *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.* **2009**, *31* (4), 671−690.
(31) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, 1st ed.; Clarendon Press: Oxford, 1991; DOI: 10.1093/oso/9780198803195.001.0001.
(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*, bau116.
(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. Cheminf.* **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) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *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; DOI: 10.1201/9780203903827.ch5.
(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. *Biol. Chem.* **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* (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.

https://doi.org/10.1021/acs.jctc.1c00254
J. Chem. Theory Comput. 2021, 17, 3554−3570

CONCLUSION

The goal of this work was to extend the CHAMM FF and CGenFF to a large set of 333 nonstandard amino acids frequently present in PDB structures and SwissSidechain database. The nonstandard amino acids were selected to be of both natural and artificial origin and to present chemical modifications at the level of sidechain and/or at the level of the backbone group. Force field parameters including partial charges, bond, valence angle, dihedral and improper torsion terms were considered for optimization. The optimization was performed according to the standard CHARMM method to ensure the compatibility between nonstandard amino acids force field developed in this work and with other components of the simulation system, described by the standard CHARMM FF. The validation of the developed force field was achieved through MD simulations of protein systems containing nonstandard amino acids. The protein models in the MD simulations mostly fluctuated around the experimental structures as demonstrated by different criteria.

The main drawback of the current implementation is that the backbone modified nonstandard amino acids are represented as a mixture of atom types and corresponding parameters from both C36 and CGenFF. Many parameters overlap in C36 and CGenFF, and in future these FFs will be united, using uniform atom types for standard, nonstandard amino acids, and small molecules. However, the current implementation is efficient and can be used to study various molecular interactions systems. One such study is present in **Chapter 4** to predict the interaction between a phenylalanyl group covalently bond to tRNA and its partner, a cyclodipeptide synthase.

During the FF development for the nonstandard amino acids, we observed that the empirical model not always reproduces the QM geometry of molecules in vacuum, which can be due to the limited FF functional form. As we demonstrated, while not being important for applications in a general case, this structural deviation between QM and FF structures leads to suboptimal force constants in the FF development. To address this issue, in **Chapter 5** we developed a new method for the FF development.

Chapter 4
FF DEVELOPMENT TO STUDY MODIFIED tRNA AND ITS INTERACTION WITH THE PROTEIN

Transfer RNA (tRNA) is frequently chemically modified (by methylation, and other types of modifications), and during it function, an amino acid becomes covalently bonded to the 3′-hydroxyl group on the CCA tail, which is catalysed by aminoacyl tRNA synthetases. In the current chapter the development and application of the CHARMM force field for an aminoacyl group linked to aa-tRNA is presented, which was reported in our previous publication.[103] Precisely, the interaction of a cyclodipeptide synthase (CDPS) and its substrate, an aminoacylated tRNA, is structurally characterized by computer modelling. Simulations results were further related to biochemical experiments performed by the experimental collaborators in I2BC (Dr. Muriel Gondry).

CDPSs form a family of recently-discovered enzymes catalysing the formation of cyclodipeptides via a sequential ping-pong mechanism using two aminoacyl-tRNA substrates.[131,132] CDPSs were structurally characterized in tRNA-free forms,[133–136] but there was no available structural detail of the CDPS:tRNA complex at the moment of publication. In this work, using AlbC CDPS from *Streptomyces noursei* that mainly produces cyclo(L-Phe-L-Phe) as model system, the interaction between the CDPS with its Phe-tRNA^{Phe} was investigated by a range of simulation techniques.

Ten initial binding poses of AlbC:tRNA were produced by rigid body docking of AlbC on the tRNA molecule. The models were further refined using MD simulations and ranked by binding free energy calculations leading to the creation of a final model. This model is characterized by multiple interactions between AlbC and tRNA and is stable over the simulation time. In this model, the α4 helix is positioned inside the major groove of the acceptor stem of tRNA with α4 positively charged residues interacting with the phosphate groups of tRNA. The residue component analysis of the binding free energy identified residues contributing to the binding affinity in a very good agreement with available biochemical data[135] and the results of *in vivo* assay experiments performed in this work by the experimental

collaborators. The proposed model of the complex is also compatible with the available experimental structure of AlbC in the dipeptide intermediate state[137].

The details of FF development for the aminoacyl group of tRNA can be found in Appendix. The paper describing the application of the developed FF model is included to the end of this chapter.

Biochemistry

pubs.acs.org/biochemistry Article

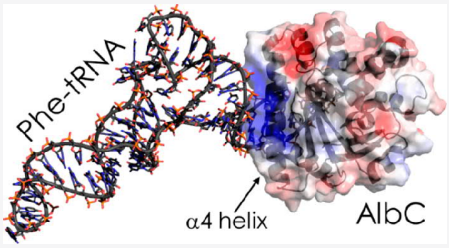
Cyclodipeptide Synthases of the NYH Subfamily Recognize tRNA Using an α-Helix Enriched with Positive Residues

Anastasia Croitoru, Morgan Babin, Hannu Myllykallio, Muriel Gondry, and Alexey Aleksandrov*

Cite This: https://dx.doi.org/10.1021/acs.biochem.0c00761 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Cyclodipeptide synthases (CDPSs) perform nonribosomal protein synthesis using two aminoacyl-tRNA substrates to produce cyclodipeptides. At present, there are no structural details of the CDPS:tRNA interaction available. Using AlbC, a CDPS that produces cyclo(L-Phe-L-Phe), the interaction between AlbC and its Phe-tRNA substrate was investigated. Simulations of models of the AlbC:tRNA complex, proposed by rigid-body docking or homology modeling, demonstrated that interactions with residues of an AlbC α-helix, α4, significantly contribute to the free energy of binding of AlbC to tRNA. Individual residue contributions to the tRNA binding free energy of the discovered binding mode explain well the available biochemical data, and the results of *in vivo* assay experiments performed in this work and guided by simulations. In molecular dynamics simulations, the phenylalanyl group predominantly occupied the two positions observed in the experimental structure of AlbC in the dipeptide intermediate state, suggesting that tRNAs of the first and second substrates interact with AlbC in a similar manner. Overall, given the high degree of sequence and structural similarity among the members of the CDPS NYH protein subfamily, the mechanism of the protein:tRNA interaction is expected to be pertinent to a wide range of proteins interacting with tRNA.



Cyclodipeptide synthases (CDPSs) make up a family of enzymes that use two aminoacyl-tRNAs (aa-tRNAs) to synthesize cyclodipeptides.[1,2] Cyclodipeptides belong to the diketopiperazine family of secondary metabolites produced in many bacteria. This class of compounds is of broad interest due to its wide range of biological activities, including antibacterial, antifungal, antiviral, antiprion, antitumor, and immunosuppressive functions.[3−10] AlbC was the first member of the CDPS family identified during the characterization of the biosynthetic pathway of the antibacterial cyclodipeptide albonoursin in *Streptomyces noursei*.[11] It mainly catalyzes the formation of cyclo(L-Phe-L-Phe) (cFF) by incorporating two phenylalanines, and cyclo(L-Phe-L-Leu) (cFL) from phenylalanine and leucine, depending on the substrate availability.[11,12]

To date, seven CDPSs, listed in Table S1, have been structurally characterized in tRNA free forms.[13−16] These CDPSs contain a Rossmann-fold domain and exist as monomers, except *Nbra*-CDPS, which is supposedly a homodimer in solution.[14] Interestingly, two residue sequences are mainly found in CDPSs at positions 40, 202, and 203 (AlbC numbering), NYH and XYP, which allowed CDPSs to be divided into two subfamilies based on residue types at these positions. AlbC belongs to the NYH subfamily. In the XYP subfamily, a proline substitutes for the histidine beside the conserved tyrosine and a nonconserved residue is found at the asparagine position.[17] NYH members incorporate 17 different amino acids into cyclodipeptides, except for His, Asp, and Lys; XYP members also incorporate 17 amino acids, differing from NYH members by the exclusion of Arg.[18,19] Structural analysis of the two subfamilies demonstrated that the main difference is in the first half of their Rossmann fold, but the catalytic residues are identical in the two families and adopt similar positions. Moreover, mutations of these residues have a similar effect on the function in both CDPS subfamilies, strongly suggesting that CDPSs of the two subfamilies share the same catalytic mechanism. Thus, it was proposed that the XYP and NYH motifs appeared as alternative solutions to the same enzymatic problem but were adopted for differences in the Rossmann fold in the two subfamilies.[14]

The catalytic mechanism has been extensively studied experimentally for the structurally characterized CDPSs.[13,15,16,20,21] In particular, AlbC was structurally characterized with a covalently attached dipeptide analogue corresponding to the reaction intermediate state before the final cyclization step.[20,21] The catalytic cycle begins with the binding of the first aa-tRNA with the aminoacyl group accommodated in the deep and hydrophobic P1 pocket that contains the conserved catalytic residues.[15] The subsequent transfer of the aminoacyl moiety to the conserved serine

Received: September 14, 2020
Revised: December 1, 2020

residue, Ser37, leads to the formation of an aminoacyl enzyme intermediate. For the second step, the $tRNA^{Phe}$ part of the first substrate dissociates from AlbC and a second aa-tRNA binds to the enzyme with its aminoacyl group accommodated in the wider P2 cavity close to the P1 pocket. The phenylalanyl-AlbC reacts with the second aa-tRNA to form a dipeptidyl-AlbC intermediate.[15,20] In the last step, the cyclodipeptide product is obtained through intramolecular cyclization. Residues important for the reaction in AlbC were identified through site-directed mutagenesis and biochemical studies.[15,20] These residues including Ser37, Tyr202, Tyr178, Glu182, Asn40, and His203. Tyr178 and Glu182 are involved in the stabilization of the aminoacyl moiety (named Phe1) of the first phenylalanyl-tRNA throughout the catalytic cycle as suggested by the crystal structure of the diphenylalanyl-enzyme intermediate mimic. The hydroxyl group of Tyr202 serves as a proton relay in the last step of the cyclization reaction, as demonstrated recently by computer modeling.[21] Three residues in AlbC, Asn40, Tyr178, and His203, help to maintain the correct reactive conformation of the dipeptidyl group[20] during the last cyclization step.

The binding of AlbC to the first tRNA appears to be contributed by interactions of the substrate aminoacyl moiety in pocket P1 and interactions of the tRNA moiety with the patch of basic residues on helix α4.[15] Residues Arg80, Arg91, Lys94, Arg98, Arg99, and Arg102, all except Arg80 belonging to α4, were identified as being important for the cyclodipeptide-synthesizing activity. Their mutation to alanine decreases considerably the level of production *in vivo* and *in vitro*.[15,20] The AlbC specificity for the first substrate is also contributed by interactions with the aminoacyl moiety, and not at the sequence of the tRNA. However, AlbC seems to handle differently its second substrate. In particular, Asn159, Arg160, and Asp163 of the α6−α7 loop and Asp205 of the β6−α8 loop of AlbC together with the aminoacyl moiety and the G^1-C^{72} base pair of the acceptor arm appear to be important for interactions with the second aa-tRNA substrate.[12] CDPSs are structurally similar to the catalytic domains of class Ic aminoacyl tRNA synthetases (aaRSs), suggesting that CDPSs probably evolved from these aaRSs[15,22] or from a common ancestor. In class Ic aaRSs, the loops equivalent to the AlbC α6−α7 loop are known to be implicated in tRNA binding,[13] which made Moutiez et al. suggest that the interaction of the second substrate with AlbC could be similar to that observed for tRNA binding to class Ic aaRSs.

Nevertheless, CDPSs and TyrRSs differ significantly. CDPSs do not have the C-terminal domain present in aaRSs and needed to recognize the anticodon and also lack the ATP binding motifs, because there is no need to activate amino acids. Two other protein families, FemX aminoacyl-transferases and aa-tRNA protein transferases, bind aa-tRNAs to create peptide bonds; however, they are structurally different from CDPSs because they possess a GCN5-related *N*-acetyltransferase (GNAT) fold.[2,23,24] FemX from *Weissella viridescens*, widely used as a model transferase for experimental purposes, catalyzes the addition of alanine from $Ala\text{-}tRNA^{Ala}$ to the UDP-MurNAc-pentapeptide.[25,26] It interacts with the acceptor arm of the tRNA moiety, as it is still able to interact with an artificial helix mimicking the acceptor arm of tRNA. Similar to CDPSs, the G^2-C^{71} base pair was shown to be important for the tRNA recognition in addition to the aminoacyl moiety.[23] In the second protein family, leucyl/phenylalanyl-tRNA protein transferase (L/F transferase) catalyzes peptide bond formation by using $Leu\text{-}tRNA^{Leu}$ (or $Phe\text{-}tRNA^{Phe}$) as a donor substrate and its terminal Arg (or Lys) as an acceptor substrate. Biochemical and structural studies indicate that L/F transferase interacts with two sequence regions in the acceptor stem of tRNA, the G^3-C^{70} base pair and a set of four nucleotides (C^{72}, A^4-U^{69}, C^{68}). Similar to FemX, L/F transferase can efficiently interact with a helical mimic of the tRNA acceptor stem. Moreover, similar to AlbC, L/F transferase has a strong preference for the CAG isoacceptor of $Leu\text{-}tRNA^{Leu}$.[12,24,27] Finally, a recent crystallographic and biochemical study showed that the CDPS from *Candidatus Glomeribacter gigasporarum* belonging to the XYP subfamily interacts with the major groove of the acceptor stem of tRNA through the basic residues of strands β2 and β7.[14] However, these residues, as demonstrated by biochemical experiments, are not implicated in interactions with tRNA in CPDPs of the NYH subfamily, demonstrating that CDPSs of the two subfamilies do not share the same mode of interaction with tRNA.[12,15]

Here we show for the first time structural details of how NYH CDPSs interact with the tRNA moiety of their substrates. We studied the interaction between AlbC and its Phe-tRNA substrate using computational techniques, including molecular dynamics (MD) simulations and binding free energy calculations. We propose a model that explains previous biochemical and structural experiments. On the basis of this model, mutagenesis experiments were performed in this work that further corroborate the model. In particular, two residues untested previously were tested in *in vivo* assays, demonstrating effects in agreement with the binding contributions of these residues in the model. In this model, the acceptor stem of the tRNA substrate interacts with the basic residues of an α-helix, α4, present in all CDPSs, while the phenylalanyl moiety predominantly occupies the two positions previously identified for the first and second substrate. The total charge of the α-helix is well-conserved in enzymes of the NYH subfamily, suggesting that the binding mode of tRNA is shared by all NYH CDPSs. Moreover, residues that were proposed to be important for binding of the first and second tRNA substrates were found to interact with tRNA in this model, strongly suggesting that the first and second tRNA substrates both interact with helix α4 in a similar binding mode.

METHODS

Residue and Charge Conservation Analysis. The sequences of CDPSs with a level of sequence identity of <90% were selected for analysis from ref 18. These sequences were further divided into two groups according to the signature sequences NYH and XYP identified previously.[17] Structural alignments were used to guide the sequence alignment with Clustal Omega software[28] for each subfamily. For the NYH subfamily, AlbC [Protein Data Bank (PDB) entry 4Q24[20]], Rv2275 (PDB entry 2X9Q[16]), and YvmC (PDB entry 3OQH[13]) were used. For the XYP subfamily, *Nbra*-CDPS (PDB entry 5MLQ[14]), *Rgry*-CDPS (PDB entry 5MLP[14]), and *Fdum*-CDPS (PDB entry 5OCD[14]) were used. The structural alignment was performed with the PyMOL software.[29] To understand better the residue conservation in the context of interactions with tRNA, the average total charge of different protein regions was calculated as follows. From the NYH alignment, the region corresponding to helix α4 in AlbC (residues 81−106) was considered for each sequence and the total charge of the region was calculated by subtracting the

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

number of negatively charged residues, aspartates and glutamates, from the number of positive residues, lysines and arginines. Histidines were assumed to be neutral. In the XYP alignment, the region corresponding to helix α4 in *Nbra*-CDPS (residues 78−102) was used. A similar approach was used for helices α5 and α6.

AlbC:tRNA Model Building. The structure of $tRNA^{Phe}$ was retrieved from the PDB, entry 4YCO (chain D), corresponding to dihydrouridine synthase in complex with $tRNA^{Phe}$ from *Escherichia coli* with resolution of 2.1 Å.[30] The chosen $tRNA^{Phe}$ structure is one of the most complete $tRNA^{Phe}$ structures available in the PDB and contains 74 residues of 76 with defined coordinates. This $tRNA^{Phe}$ structure has no post-transcriptional modifications; however, it is similar to the mature tRNA and can be efficiently used by AlbC as demonstrated previously.[12] This tRNA contains two trans-versions, C3-G70 and G3-C70, in the acceptor stem, which were modeled as in the crystal structure.[31] The structure of 4YCO was obtained at very high concentrations of Mg ions (200 mM $MgCl_2$),[30] explaining the high level of magnesium ions found in the crystal structure (25 Mg ions total). Thus, of eight magnesium ions present in chain D of the crystallographic structure, four magnesium ions interacting with phosphate groups of tRNA were maintained in the model. The other four magnesium ions were not considered, because they interact with fewer than two phosphates of the tRNA backbone, and were not observed in the other crystal structure of $tRNA^{Phe}$ (PDB entry 1I9V).[32] The aminoacyl group and missing residues 75 and 76 were initially built in the extended conformation and were energetically minimized with the rest of the tRNA structure fixed. The CHARMM36 force field was used for the protein[33,34] and the TIP3P model for water.[35−37] The phenylalanyl group was modeled using the force field model specifically developed as a part of this work. The details of the force field development as well as the force field model are given in the Supporting Information.

The crystal structure of AlbC, PDB entry 4Q24,[20] was used for the AlbC model. As previously described,[21] the structure of wild-type AlbC was built from the existing crystal structure by converting Sγ of Cys37 into an oxygen and deleting the ZPK ligand (*N*-carbobenzyloxy-L-Phe-methyl ketone). The same protein protonation state was assigned as in the previous study,[21] except Glu182, which was in the deprotonated form. This glutamate acts as a catalytic base and becomes protonated only after the reaction with the first aa-tRNA.[21]

To acquire models of the AlbC:tRNA interaction, rigid-body docking was applied with the ZDOCK server,[38] proven to be a valuable tool for protein:RNA complex docking.[39] The docking of tRNA to AlbC was performed with ZDOCK 3.0.3 default parameters using the coordinates of AlbC and Phe-$tRNA^{Phe}$. The results from the top 2000 ZDOCK predictions were filtered using the condition that amino-acylated adenosine of tRNA interacts with residues 35, 37, 178, 182, and 202 of AlbC to position the aminoacyl moiety in the catalytic pocket. Predictions were kept only if all selected residues were within 6 Å of the docked tRNA and resulted in 32 models. The first 10 top score models were selected for further analysis. An additional model of the AlbC:tRNA interaction was created on the basis of the structural homology between CDPSs and synthetases as described in the Supporting Information.

Preliminary calculations on AlbC:tRNA interactions were performed using smaller models. In the small models, tRNA residues that are distant from the protein and, thus, not expected to contribute to the protein binding were excluded from simulations. In particular, tRNA nucleotides with a distance to the protein of >26 Å were deleted; hence, residues 23−45 of phe-$tRNA^{Phe}$ were not considered.

Molecular Dynamics Simulations. MD simulations were performed with the NAMD simulation package.[40] Each protein:tRNA complex was immersed in a box of water, the sides of which were at least 12 Å from any atom of the tRNA protein. Water molecules overlapping the protein and tRNA were removed. Periodic boundary conditions were applied, and the entire box was replicated periodically in all directions. All long-range electrostatic interactions were computed efficiently by the particle mesh Ewald method,[41] while the short-range nonbonded interactions were calculated with a cutoff of 11 Å. An appropriate number of potassium counterions was included to render the system electrically neutral. MD simulations were performed at constant room temperature and pressure, after thermalization for 31 ps. The CHARMM36 force field[34,42,43] was used for the protein, tRNA, and the modified version of the TIP3P water model.[35]

For the small models, harmonic restraints with a force constant of 1 kcal mol^{-1} $Å^{-2}$ were applied on heavy atoms within 6 Å of the truncation region using the tRNA crystal structure as a reference. Initially, MD simulations were performed for 30 ns on each model and continued for at least 200 ns for three models (9, 2, and 7) characterized by the strongest interactions between the protein and tRNA.

The model with the complete tRNA and protein molecules was built on the basis of the tRNA position observed in model 7 characterized by the lowest binding free energy among the models. The crystal structure of PDB entry 4Q24[20] was superposed on the snapshot of model 7 with the lowest protein:tRNA binding free energy using the protein backbone atoms, and the coordinates of tRNA residues were retained. The coordinates of the nucleotides missing in the truncated model were taken from the complete $tRNA^{Phe}$ crystal structure of PDB entry 4YCO.[44] However, the tRNA aminoacyl group in model 7 was misoriented relative to the experimental structure (PDB entry 4Q24), as shown in Figure S1, suggesting that MD simulations of model 7 have not converged. To correct the position of the aminoacyl group in model 7, we used its position in model 9, which is close to the experimental position in the crystal structure (as demonstrated in Figure S1) as follows. The tRNA was superimposed on the lowest-energy structure of model 9, and tRNA terminal residues 75 and 76 with the aminoacyl group were retained for the final model. The model was then energetically minimized using 200 minimization steps in the CHARMM software with restraints applied on residues 75 and 76 of the nucleic acid.[45] The model with the complete tRNA contained around 162000 atoms, and the small models contained around 127000 atoms each. For the model with the complete tRNA molecule, a 1 μs MD simulation was performed with the center of mass of the protein and tRNA heavy atoms weakly restrained to the origin of the system by a harmonic potential with a force constant of 0.1 kcal mol^{-1} $Å^{-2}$ to prevent the drift of the AlbC:tRNA complex in MD simulations.

Binding Free Energy Calculations. The free energy of binding of Phe-$tRNA^{Phe}$ to the AlbC protein was estimated as the difference in the total free energy of the complex and the free energies of the separated partners:

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

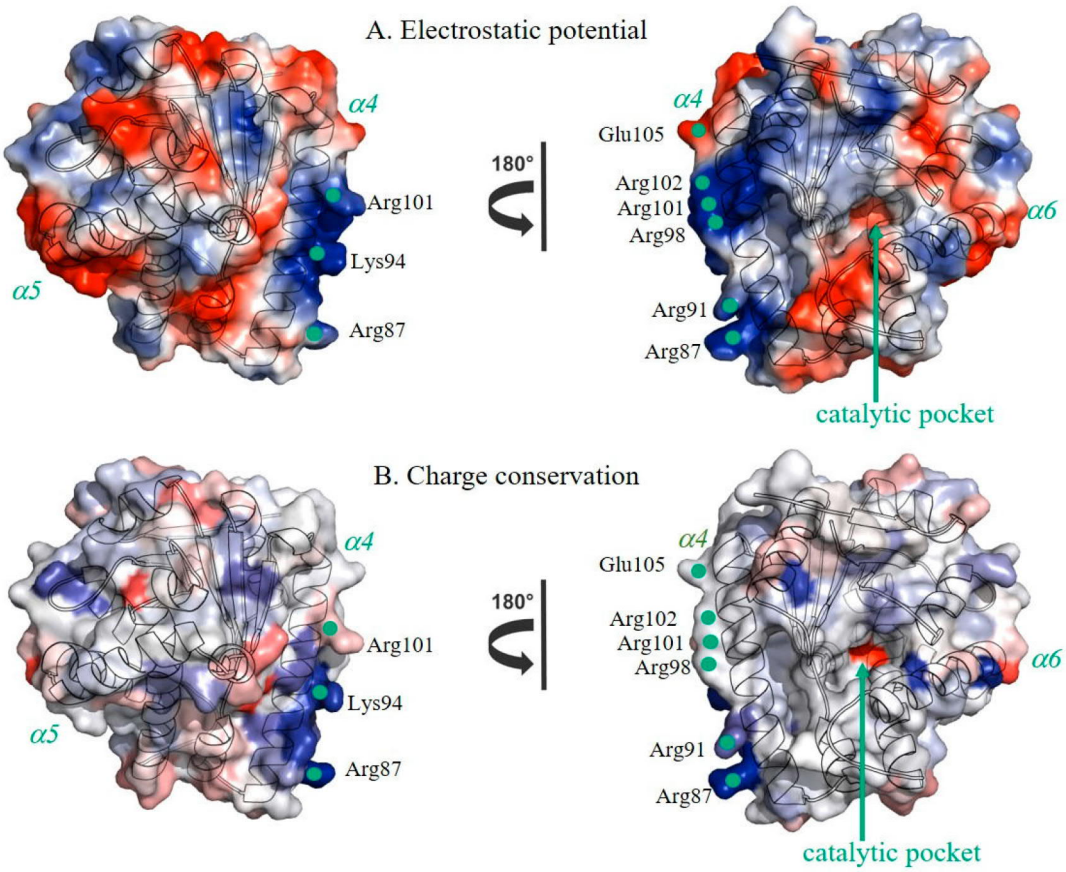


Figure 1. (A) Electrostatic potential on the AlbC surface and (B) conservation of the residue charge in the NYH subfamily of CDPSs. Green dots show the centers of mass of the important indicated residues. (A) Positive and negative potentials are colored blue and red, respectively. (B) Average residue charge obtained from the sequence alignment and considering lysines and arginines positively charged and aspartates and glutamates negatively charged.

$$\Delta G_{\text{prot:tRNA}} = G_{\text{prot:tRNA}} - G_{\text{prot}} - G_{\text{tRNA}} \quad (1)$$

The free energy (G) has three contributions from polar interactions computed using the Poisson−Boltzmann model (G_{PB}), the nonpolar (G_{SA}) term, and the vibrational entropy:

$$G = G_{\text{PB}} + G_{\text{SA}} - TS \quad (2)$$

where T is room temperature. The Poisson−Boltzmann equation was solved numerically with the PBEQ module[46,47] implemented in CHARMM version c41b1.[45] The dielectric constants for the solute and solvent volumes were set to 4 and 80, respectively. The solute−solvent dielectric boundary was defined as a molecular surface using a water probe with a radius of 1.4 Å. The Poisson−Boltzmann equation was solved using a cubic grid and a finite difference algorithm. A two-step protocol was used with an initial calculation performed using a large box with a coarse grid providing the boundary conditions for a second calculation with a smaller box using a finer grid.[48] The coarse grid spacing and fine grid spacing were set to 0.8 and 0.4 Å, respectively, and the grid size was chosen to include the entire protein:tRNA complex in both calculations. A physiological ionic strength with a monovalent ion concentration of 0.15 M was used in addition to the four structural magnesium ions. The nonpolar contribution was estimated by the term proportional to the solvent accessible surface area (SASA):

$$G_{\text{SA}} = \alpha \times \text{SASA} \quad (3)$$

where the surface tension α = 6 cal mol^{-1} Å^{-2}.[49]

Separate tRNA and protein structures were obtained by simply discarding the unwanted partner. Thus, structural relaxation upon dissociation was not explicitly modeled but included implicitly in the higher protein internal dielectric constant.[50,51] Calculations were performed for 300 snapshots taken each 100 ps from the 30 ns MD simulations for each small model. For models 9, 2, and 7, binding free energy calculations were performed for at least 2000 snapshots taken each 100 ps from the MD simulations. For the model with the complete tRNA molecule, free energy calculations were performed on 1000 structures taken each 1 ns from the 1 μs MD simulation.

The conformational entropy was estimated by normal mode analysis (NMA)[52,53] on 10 snapshots taken each nanosecond during the last 10 ns of MD simulations. To calculate normal modes, all water molecules were removed and the system containing the protein and tRNA was energetically minimized using the Adopted Basis Newton−Raphson minimizer implemented in CHARMM.[45] The distant-dependent dielectric constant of four was used in this calculations. The tolerance applied to the average gradient of 0.0001 was used as the convergence criterion. The error of the vibrational entropy calculations was estimated by dividing energies corresponding

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

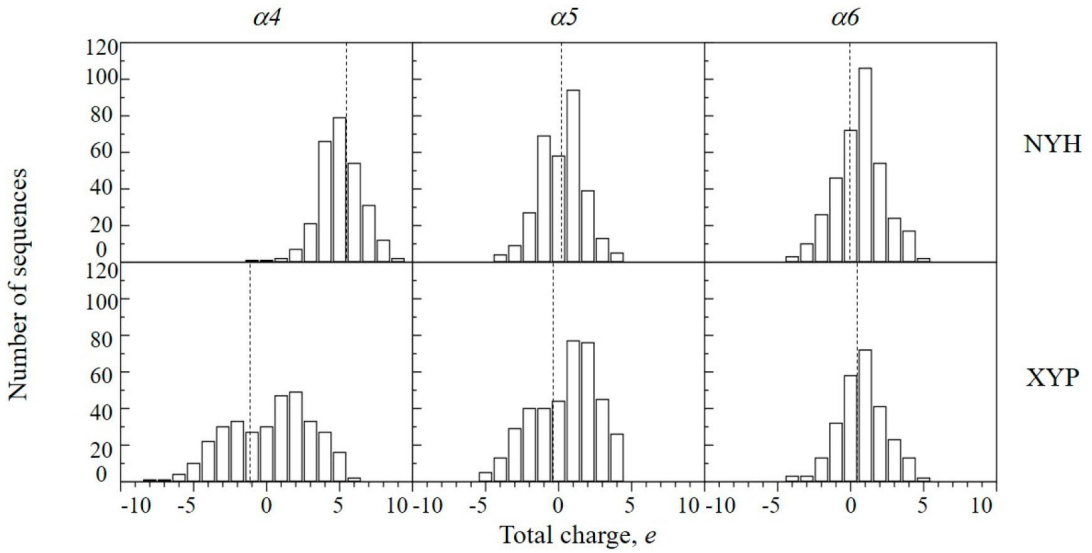


Figure 2. Distribution of the total charge of α-helices α4−α6 in the NYH and XYP subfamilies of CDPSs obtained after the sequence alignment. The dashed line represents the mean charge.

to the 10 snapshots into two batches and computing the difference.

To estimate long-range electrostatic effects of residue Asp95 in the catalytic center, the Poisson−Boltzmann (PB) model was used.[54] The electrostatic potential on atoms of the catalytic residues was computed using the PB model and averaged over structures taken from 100 ns MD simulations with the protein in the dipeptide−intermediate state. The same setup described above was used for Poisson−Boltzmann calculations. Calculations were repeated with zero charges on the Asp95 side chain, and the difference in the electrostatic potential was calculated to obtain the electrostatic potential due to the charge on Asp95 and the associated solvent response.

Individual Residue Contribution to the Binding Free Energy. The individual contribution of protein residues to protein:tRNA binding was estimated by component analysis, as described previously.[55,56] Only the electrostatic contribution was calculated, as it is expected to be dominant in interactions between the protein and tRNA. Charges on a side chain up to Cβ of the residue were zeroed during binding free energy calculations. The energy contribution of the residue is calculated as the difference between the binding free energy of the wild-type complex and the energy of the complex with zero charges on the residue side chain.

***In Vivo* Assays for Wild-Type AlbC and Lys46Ala and Asp95Ala Variants.** The expression plasmid encoding AlbC was previously constructed.[15] The plasmids encoding AlbC variants were generated by PCR mutagenesis according to the QuikChange method (Stratagene). The plasmids were used to transform *E. coli* BL21AI [pREP4] cells. The strains were grown at 37 °C in the M9 minimum medium supplemented with trace elements and vitamins, 0.5% glycerol (0.5% glucose for starter cultures), 200 μg/mL ampicillin, and 25 μg/mL kanamycin. CDPS expression was induced by isopropyl β-D-thiogalactopyranoside (IPTG, final concentration of 2 mM), and cultivations were continued for 18 h at 20 °C. The cultures were centrifuged at 4000*g* for 45 min: supernatants and cell pellets were analyzed for cyclodipeptide-synthesizing activities and protein expression, respectively. Supernatants were acidified with trifluoroacetic acid at a final concentration of 2% (v:v) and submitted to LC-MS analyses as previously described.[57] Cell pellets were frozen at −80 °C and broken as described in ref 1. The soluble protein fractions were separated from the insoluble fractions by centrifugation at 20000*g* for 45 min. Both fractions were analyzed by 12% sodium dodecyl sulfate−polyacrylamide gel electrophoresis with Coomassie blue staining.

RESULTS

Conservation of the Positive Charge of CDPS Helix α4 Belonging to the NYH Family. The residue conservation was first analyzed in the context of interactions with tRNA. In particular, positive charges of protein residues may contribute strongly to tRNA binding. The electrostatic potential on the AlbC surface calculated using the experimental structure of PDB entry 4Q24[20] is shown in Figure 1. The total charge of AlbC at pH 8 is −4 *e*; however, the protein is strongly polarized at this pH, which is manifested in relatively large patches of positive and negative electrostatic potential on the protein surface in Figure 1. It was shown that proteins with such strong polarization are implicated in binding to the ribosome or interact with nucleic acids.[58] In particular, a significant region of positive potential is found on helix α4. Helix α4 is comprised of 26 residues, with 13 residues being ionized at pH 8, including seven arginines, two lysines, three glutamates, and one aspartate. The total charge of helix α4 at pH 8 is, thus, +5 *e*. The potential on the protein surface, particularly in the catalytic pocket and around helix α4, correlates well with the elevated level of charge conservation in the NYH subfamily of CDPSs (Figure 1B).

The residue charge conservation shown in Figure 1 is clearly visible for the residues in the catalytic pocket, and in particular for Glu182, implicated in the enzymatic activity.[15] Positively charged residues of helix α4 possess a relatively significant level of conservation of the ionization state but demonstrate no strict conservation. More precisely, helix α4 is enriched in arginines and lysines in CDPSs in the entire NYH subfamily; however, these residues have scattered positions in the alignments published previously.[1,14,15] Thus, in the NYH

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

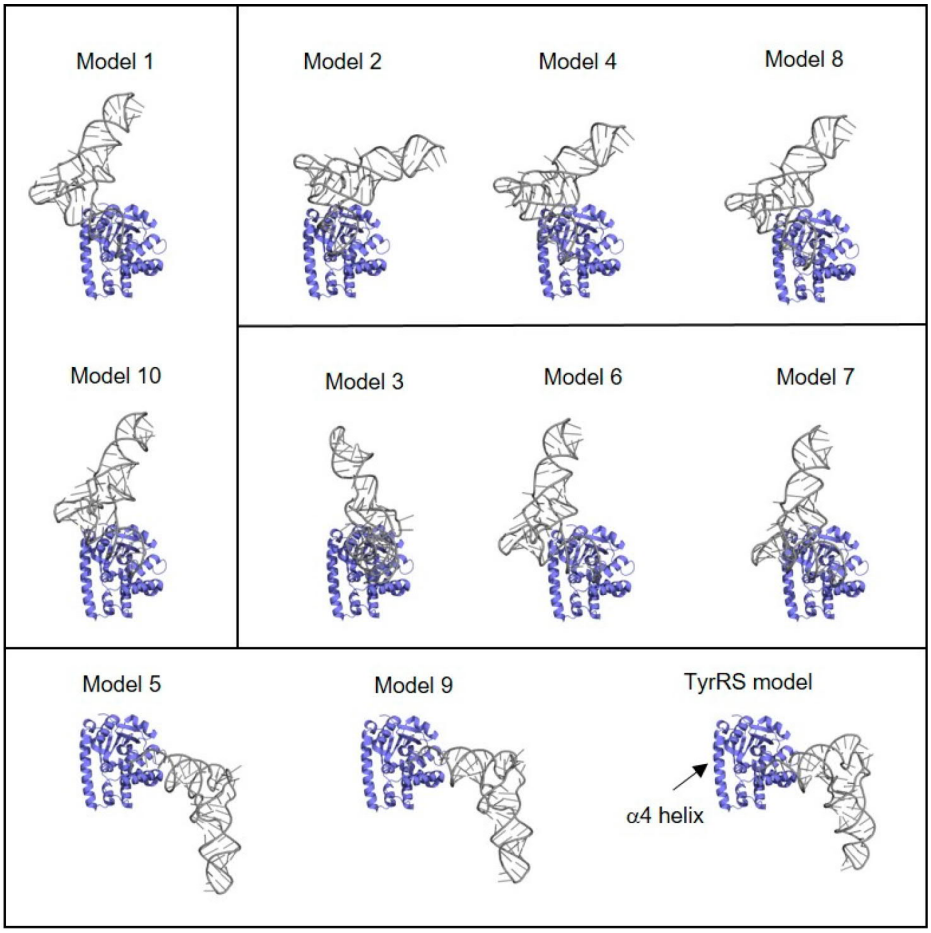


Figure 3. Initial models of the AlbC:Phe-tRNA^{Phe} complex derived by docking and homology modeling. Ten models were obtained by rigid-body docking, and one additional model was based on the experimental structure of the TyrRS:tRNA complex. The models were superimposed using the protein atoms and classified according to structural similarity. The model number corresponds to their docking scoring rank.

subfamily (272 sequences), the total positive charge of helix α4 is conserved with an average charge of +5.5 e and a standard deviation (SD) of 1.5 e as shown in Figure 2. This suggests that the positive residues of helix α4 could be implicated in tRNA binding. In contrast to other solvent-exposed helices, helix α4 is the longest helix and the only helix having significant positive charge. Other helices on the protein surface, α5 and α6, have average total charges of 0.3 e (SD of 1.7 e) and −0.2 e (SD of 2.2 e), respectively (Figure 2). Interestingly, in the XYP subfamily (231 sequences), the charge of helix α4 is small and slightly negative with an average of -1.2 ± 2.6 e, while helices α5 and α6 are almost neutral, suggesting that helix α4 may accomplish different functions in tRNA binding in these two subfamilies.

Rigid-Body Docking Study of AlbC:tRNA Interactions. To design initial models of the AlbC:tRNA complex, rigid-body docking of the first tRNA substrate to AlbC was performed with the ZDOCK server.[38] The models of the docked complex were filtered on the basis of the proximity of the important residues in the catalytic site to tRNA as described in Methods. However, the proximity requirement with the minimal distance set to 6 Å was not strict, and no filtering restraints were applied to interactions with other protein residues, including residues of helix α4. Thus, rigid-body docking provided 10 initial models for the AlbC:tRNA complex. One additional model for the first substrate binding was designed on the basis of the structural similarity between AlbC and tyrosyl-tRNA synthetase.[15] Eleven models were then classified on the basis of the root-mean-square deviation (RMSD) (Table S2) giving four groups of models shown in Figure 3. Among the 11 considered models, only one subgroup of models 5 and 9 and the model based on TyrRS presented no interactions with helix α4.[15] This group of models cannot explain the experimental evidence of the contribution of residues of helix α4 to binding of the first tRNA substrate. Nevertheless, these models were also investigated to elucidate the source of the difference between interactions of CDPS and TyrRS with tRNA substrates and to probe a hypothesis that helix α4 is involved in binding of the second tRNA substrate despite what was previously suggested.[12] Additionally, the Poisson−Boltzmann binding free energy was computed for 2000 structures generated with ZDOCK after structural relaxation as described in the Supporting Information. The results show that the binding free energy for strongly interacting AlbC and tRNA correlates well with the contribution of helix α4, as shown in Figure S2, and the five models with the lowest binding free energy are all similar and also similar to model 7, shown in Figure S2 and Figure 3.

MD simulations were performed for each of 11 models in the explicit solvent. In preliminary calculations on the docked structures, Phe-tRNA^{Phe} residues that are at least 26 Å from AlbC were not present, as these distant residues do not

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

contribute to AlbC binding. Indeed, in these calculations, the contribution of distant groups is approximated as the contribution of the solvent occupying their space. We estimated this contribution of the truncated region using the Poisson−Boltzmann model as described in Methods and found that it contributes <0.1 kcal mol^{-1} in all cases in agreement with the previous studies.[59] To maintain the tRNA structure near the truncated region, harmonic restraints were applied on heavy atoms around the truncated region, and initially MD simulations for 30 ns were performed to evaluate the free energy of binding between AlbC and tRNA. In the MD simulations, tRNA remained positioned close to the protein and the aminoacyl group remained in the AlbC catalytic pocket in all models. Similar to their starting poses, the TyrRS-based model and models 5 and 9 did not interact with helix α4 during the simulation, while in the remaining models, tRNA preferred to interact with helix α4. To evaluate the binding free energy, Poisson−Boltzmann/surface area (PB/SA) free energy calculations were performed using the structures drawn from the MD simulations. The vibrational entropy was estimated using normal mode analysis. The results are summarized in Table S3. Model 7 presented the lowest binding free energy, i.e., with the strongest interaction between the protein and tRNA as shown in Figure S3, and the entropy estimate decreased further the total binding free energy for model 7 relative to those of other models. Model 2 demonstrated a lower binding free energy toward the end of the simulation, suggesting that the MD simulation has not converged for this model with 30 ns. Model 9 was included in subsequent analysis because it represents the TyrRS binding pose. Thus, on the basis of the preliminary calculations, MD simulations were continued for representative models 2, 7, and 9: 300 ns for models 2 and 7 and 200 ns for model 9. The binding free energy and RMSD shown in Figure S4 did not demonstrate large fluctuations after simulations for 20 ns for models 7 and 9, suggesting that these simulations converged. The average PB/SA binding free energies observed for models 2 and 7 in MD simulations were −41.0 kcal mol^{-1} (SD of 7 kcal mol^{-1}) and −47.0 kcal mol^{-1} (SD of 7 kcal mol^{-1}), respectively. The average binding free energy for model 9 is −20.5 kcal mol^{-1} (SD of 3 kcal mol^{-1}). Including the vibrational entropy estimate decreased the binding free energy for model 9 relative to models 7 and 2; however, model 7 still has a total binding free energy that is 14.2 kcal mol^{-1} lower in comparison with that of model 9. Thus, the models based on TyrRS are characterized by significantly weaker interactions with AlbC due to the lack of interaction with helix α4, because helix α4 is one of structural elements providing the strongest contribution to the first tRNA binding. In contrast to NYH CDPSs, the total charge of helix α4 in TyrRS from *Thermus thermophilus* is zero,[60] suggesting that the binding mode of the NYH CDPS and TyrRS can be different.

Individual residue contributions to the total binding free energy for each model, listed in Table S4, were further correlated with the experimental data. In the four best models, Glu182 contributes favorably to aa-tRNA binding with an average of −5.7 kcal mol^{-1} through ionic interactions with the charged amino group of the phenylalanyl moiety (Figure S5). Glu182 is a conserved residue in all CDPSs and has an important role in the catalytic reaction by participating in the reaction and maintaining the correct orientation of the reactant groups.[15,21] In both models 2 and 7, residues Lys94, Arg98, Arg99, and Arg102 contribute strongly to the binding free energy, while Lys46 contributes mostly in model 2 and Arg91 in model 7. Biochemical experiments indicate that Arg91, Arg98, Arg99, and Arg102 are important for cyclodipeptide formation.[12,15] The mutation of Lys46 into an alanine does not have any significant effect on the cyclodipeptide production and protein expression (Table S5 and Figure S6), suggesting that this residue is not implicated in AlbC:tRNA interaction; however, a positive residue, lysine or arginine, is frequently found at this position in the CDPS NYH subfamily. In model 9 based on the TyrRS:tRNA structure, suggested previously to explain the second substrate binding, Arg231 has the strongest contribution to the total binding free energy of −8.3 kcal mol^{-1}, but residues Asp163 and Asp205 do not contribute practically to interactions with the tRNA substrate. Importantly, biochemical experiments demonstrate that the contribution of Arg231 to tRNA binding is negligible and Asp163 and Asp205 are both implicated in tRNA interactions.[2,15] Overall, this demonstrates that model 9 is unlikely for the complex with the first or second tRNA substrate.

The individual residue contributions for models 2 and 7 are very similar. In model 2, residues 94−102 of helix α4 interact with the phosphate backbone of nucleotides 65−67, while in model 7, interactions occur mainly between residues 91−103 and nucleotides 62−65, as shown in Figure S5. Thus, models 2 and 7 that are representative of two groups of structures predicted by rigid-body docking converged to similar binding modes relying on the α4:tRNA interaction.

Interactions with Helix α4 Are Maintained in Long Molecular Dynamics Simulations. On the basis of the preliminary calculations described above, a model with the complete tRNA molecule was built using the tRNA position observed in model 7, which is characterized by the strongest interaction between the protein and tRNA. In the model with the complete tRNA, the orientation of the aminoacyl group in model 7 was improved using the information available from the experimental structure (PDB entry 4Q24). In particular, the orientation in model 9, where the structure of the aminoacyl group is close to its position in the experimental structure of AlbC complexed with the dipeptide analogue, as shown in Figure S1, was used as described in Methods. The model was subjected to a long MD simulation of 1 μs with no restraints applied to the protein and tRNA atoms. The RMSD is shown in Figure S7. The average backbone RMSD for the protein referenced to the crystal structures of the protein (PDB entry 4Q24) was 1.8 Å with a standard deviation of 0.2 Å. In agreement with previous studies, tRNA demonstrates greater structural fluctuations than the protein with the backbone average RMSD of 4.0 Å (SD of 1.8 Å) referenced to the crystal structure of tRNA^{Phe} (PDB entry 4YCO). However, the position of AlbC relative to tRNA is well maintained during the entire simulation with the tRNA acceptor arm preserving the contact with helix α4 (Figure 4A). Moreover, the covariance of atomic displacements of the two partners indicates that the AlbC catalytic site and helix α4 fluctuations are correlated with those of the tRNA during the MD simulations, indicating that AlbC interacts strongly with tRNA through its catalytic site and helix α4 (Figure 4B).

The free energy of binding of AlbC to tRNA observed in MD simulations is shown in Figure S8. The binding free energy fluctuates near the average value of −40.3 kcal mol^{-1}, but fluctuations are small with a standard deviation of 5.0 kcal mol^{-1}. Overall, small variations in the binding free energy and RMSD given above suggest that MD simulations of the

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

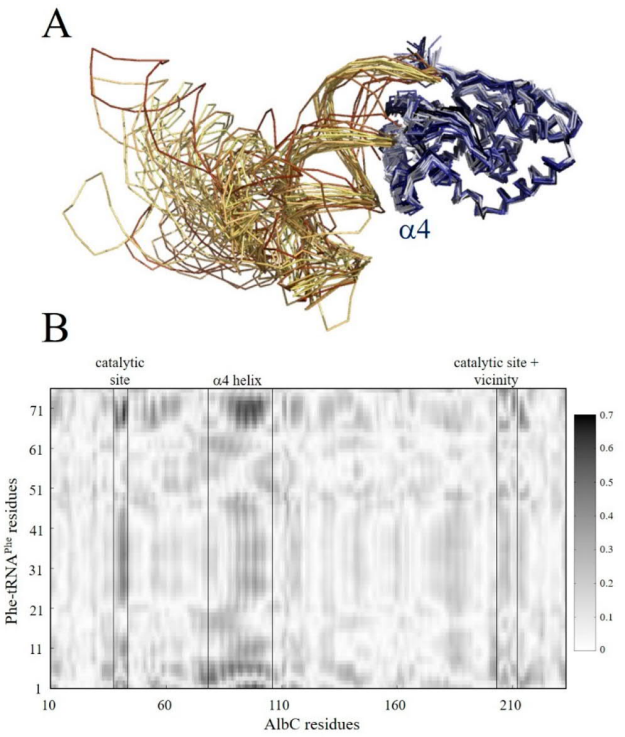


Figure 4. (A) Interactions between AlbC and tRNA observed in MD simulations of the AlbC:tRNA complex. Twenty superimposed snapshots are shown and were taken each 50 ns from 1 μs MD simulations. (B) Absolute covariance for the atomic displacements of the protein backbone Cα atoms and tRNA backbone phosphate groups. Darker areas correspond to large values of the absolute covariance.

AlbC:tRNA complex have mostly converged. The individual residue contributions to the total free energy for binding of AlbC to tRNA are listed in Table 1. Residues that are contributing to a total binding free energy of less than −7 kcal mol^{-1} (Arg91, Lys94, Arg98, and Arg99) all belong to helix α4. Positively charged residues of helix α4, Arg80, Arg87, Arg91, Lys94, Arg98, Arg99, and Arg102, create multiple salt bridges with the phosphate backbone as shown in Figure 5A. Interestingly, in MD simulations, both Arg80 and Arg99 stay very close to the tRNA substrate [5.4 and 4.0 Å, respectively (Figure 5A)] and thus can contribute through long-range electrostatic interactions. Arg80 interacts through long-range electrostatic interactions with Asp163, and Arg99 is relatively close to the C72 nucleotide both involved in the second substrate binding,[12] suggesting that the second substrate binding pose could be similar to that of the first substrate.

Table 1. Contributions of Individual Residues to the Binding Free Energy in Kilocalories per Mole[a]

residue	calculated	experimental
Lys46	−1.9 (3.0)	no effect[b]
Arg80	−2.6 (1.2)	strong
Arg87	−3.4 (1.6)	no effect
Arg91	−11.3 (3.3)	strong
Lys94	−9.2 (2.6)	strong
Asp95	4.4 (0.9)/8.9 (6.3)[c]	strong[b]
Arg98	−7.4 (2.2)	strong
Arg99	−9.0 (2.8)	strong
Arg102	−3.1 (2.1)	strong
Glu182	−6.4 (1.2)	strong
Arg231	−0.4 (0.2)	no effect

[a]Average values over 100 frames are given. The standard deviation is given in parentheses. The energies are given only for residues with the absolute contribution to the total binding free energy being >1.0 kcal mol^{-1} (except Arg231). [b]*In vivo* assay results from this study (Table S5); the experimental data for the remaining residues were compiled from previous studies.[12,15,20] [c]Results from the component analysis of MD simulations of the wild-type protein and binding free energy difference computed using MD simulations with the Asp95Ala variant are given.

Asp95 belonging to the same helix makes a positive contribution to the total binding free energy of 4.4 kcal mol^{-1}, weakening tRNA binding. To further study the role of Asp95, we performed 200 ns MD simulations of the Asp95Ala variant with the complete tRNA. The free energy of binding of the AlbC Asp95Ala variant to tRNA observed in MD simulations has small fluctuations shown in Figure S9 with an average value of −49.2 kcal mol^{-1} (SD of 6.3 kcal mol^{-1}). Compared to the wild-type MD simulations, the average binding free energy of the mutant is 9 kcal mol^{-1} lower, suggesting that Asp95 has a destabilizing role in the AlbC:tRNA interaction, due to repulsive electrostatic interactions between the Asp95 carboxylate group and tRNA phosphates. This mutation was further tested using *in vivo* experiments. The Asp95Ala variant was shown to be approximately 2−3 times less efficient than the wild-type protein (Table S5 and Figure S6), which demonstrate that Asp95 is implicated in interactions with tRNA. However, the effect of the Asp95Ala mutation found experimentally is reversed in comparison to the simulation results. We further tested if Asp95 can affect the enzymatic activity through long-range electrostatic effects. In particular, the electrostatic potential due to the charge of Asp95 was estimated as described in Methods using MD simulations of the AlbC in the dipeptide state. Asp95 is far from the catalytic center: the shortest distance in the experimental structure of PDB entry 4Q24 between the dipeptide and Asp95 (Cγ of Asp95 and C of Phe1) is 16.4 Å, and the distance between Cδ of the catalytic Glu182 and Cγ of Asp95 is 20.3 Å. In agreement with the long distance from the catalytic center and the fact that this residue is solvent-exposed, the results show that the electrostatic effects due to Asp95 in the catalytic center are very small. For example, the free energy of interaction between Asp95 and the bound dipeptide is just −0.04 (0.01) kcal mol^{-1}, while electrostatic potential due to Asp95 on the closest atom Oγ of Ser37 is just 0.09 (0.01) kcal mol^{-1} e^{-1}. Overall, the long-range electrostatic effects of Asp95 can be neglected. Finally, we tested if allosteric effects can be implicated with the Asp95Ala mutation. Figure S10 compares average structures observed in MD simulation with Asp95 and Asp95Ala with AlbC in the dipeptide state and in complex with tRNA. The positions of the catalytic residues, surface residues participating in interactions with tRNA, as well as the tRNA phosphates are all practically identical, demonstrating the absence of allosteric effects. On the basis of these insights, and the fact that the computed effect of Asp95Ala is particularly strong, we propose that the Asp95Ala mutation increases the strength of interactions with Phe-tRNA^{Phe}, but also with other tRNA forms, including tRNA without the aminoacyl group. This can have a total negative effect on the cyclodipeptide production in the cellular context, where AlbC should specifically bind to the

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

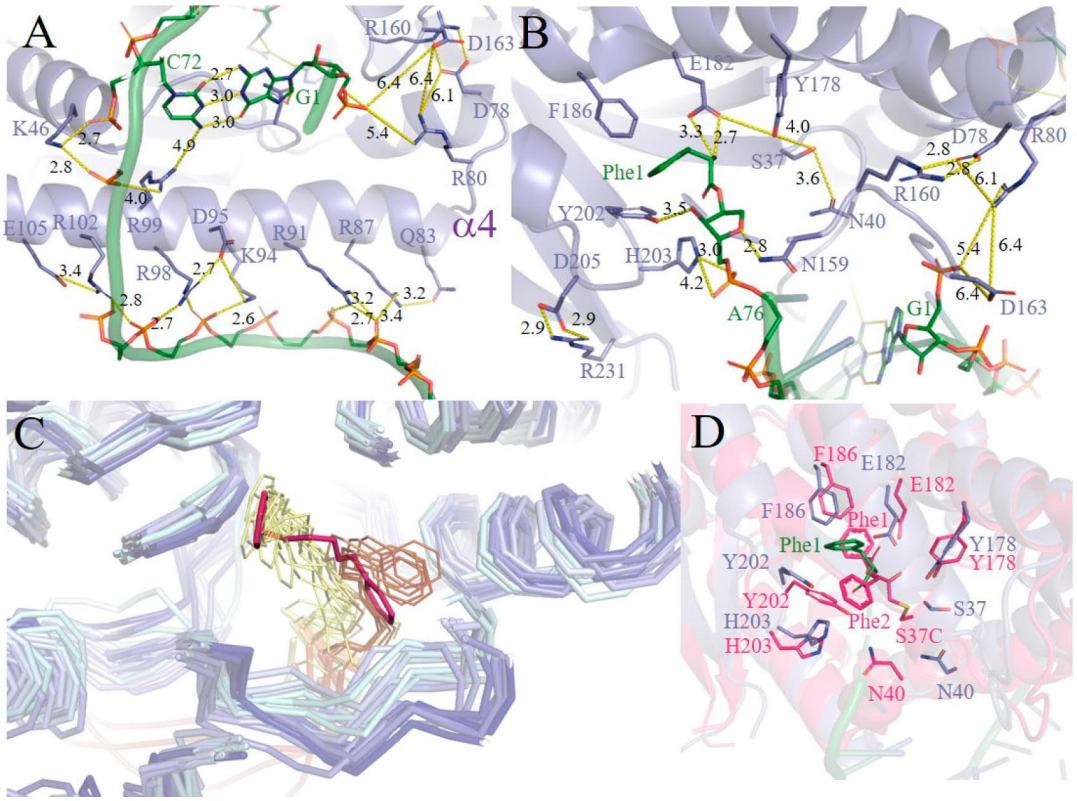


Figure 5. Interactions observed in 1 μs MD simulations of the AlbC:Phe-tRNA^{Phe} complex. The protein and Phe-tRNA^{Phe} are colored blue and green, respectively. The important interatom distances are shown in angstroms. The conformation with the binding free energy close to the average value in Figure S8 is shown. (A) Helix α4 and tRNA major groove interaction. (B) Close-up view of the catalytic center. (C) Twenty superimposed snapshots taken each 50 ns from the MD simulation demonstrating two predominant conformations of Phe1 colored yellow and orange. The structure of the dipeptide intermediate in the crystal structure (PDB entry 4Q24) is colored pink. (D) Superposition of the conformation from the MD simulation with the experimental structure of PDB entry 4Q24.[20]

correct tRNAs (Phe-tRNA^{Phe}) and dissociate from tRNA when the aminoacyl group is transferred to the protein to become available for the next catalytic cycles.

Glu182 has one of the largest contributions to tRNA binding via strong electrostatic interactions with the protonated amine group of Phe1 and helps orient and stabilize the aminoacyl in the catalytic pocket, shown in Figure 5B. Important catalytic residues Tyr202, His203, Tyr178, and Asn40 help stabilize the aminoacyl moiety through a network of hydrogen bonds. Overall, the structure of the catalytic center observed in the MD simulations is in good agreement with the crystal structure with the dipeptide substrate analogue (PDB entry 4Q24). Importantly, residues Tyr178, Glu182, and His203 essential for the catalytic function are found in a position similar to that in the crystallographic structure, shown in Figure 5D. However, two important residues, Asn40 and Ser37, are found in different orientations compared to the crystallographic structure (PDB entry 4Q24). In the crystal structure, Asn40 is solvent-exposed and close to Ser37 with a Cβ−Cβ distance of 3.9 Å. Arguably, this is an artifact of rigid-body docking, where the protein can be trapped in a higher-energy conformation, which cannot be relaxed in the presence of tRNA. Simulations of the dissociation of the protein and tRNA are complex and beyond the scope of this study. However, the effect of the misorientation of Ser37 and Asn40 is expected to be small, because these residues do not contribute significantly to tRNA binding, and thus cannot change the overall binding position of tRNA relative to the protein.

It was previously suggested[12] that residues Asn159, Arg160, Asp163, and Asp205 contribute to the second substrate binding. Importantly, the model proposed in this work can rationalize all of these contributions as shown in Figure 5B. In particular, Arg160 and Asp205 create salt bridges with Asp78 and Arg231, respectively, stabilizing α6−α7 and β6−α8 loops; Asn159 makes a hydrogen bond with the ribose of A76, and Asp163 has a repulsive electrostatic effect on the G1 phosphate group. Overall, this suggests that the binding poses for the first and second substrate may be similar. Notably, the phenylalanine group of Phe-tRNA^{Phe} occupied in MD simulations predominantly two positions, shown in Figure 5C. The two dominant orientations are compatible with the first and second phenylalanine residues of the dipeptide analogue in the crystal structure (Figure 5). This demonstrates that the first substrate can bind by its aminoacyl group into both binding pockets, P1 and P2. These pockets are occupied by the first and second phenylalanine groups in the protein crystal structure with the dipeptide analogue.

Overall, the new model presents an extended interaction between Phe-tRNA^{Phe} and AlbC through helix α4 consistent with positive charge conservation of the helix in the NYH subfamily and consistent with experimental data.

Effect of tRNA Interactions on the AlbC Conformation. Finally, to determine whether AlbC can change its conformation during its function cycle, 100 ns MD simulations were performed for AlbC in the apo and intermediate states to compare with the MD simulation of the AlbC:tRNA complex.

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

In particular, for the apo form two intendent simulations were performed using the crystal structures of PDB entries 4Q24[20] and 3OQV.[15] For AlbC in the intermediate state with Phe1 attached to Ser37, the starting structure was obtained from the structure with the dipeptide analogue by deleting Phe2 of the dipeptide analogue. The RMSD between backbone atoms of the available crystal, 4Q24, and the average structure in the MD simulations is given in Table 2.

Table 2. RMSDs Computed between the Average Structure from MD Simulations and the Crystal Structure (PDB entry 4Q24)[a]

state	RMSD (Å)
apo form modeled	1.05
apo form 4Q24	1.05
AlbC-Phe1 IS	1.05
AlbC:tRNA	1.33

[a]The RMSD was computed after the structures were superimposed on the protein Cα atoms, excluding atoms of the N- and C-termini (residues 1−14 and 231−239, respectively), due to their large structural fluctuations.

The MD simulations of the apo form starting from two different crystal structures converge to a very similar protein conformation with a 1.05 Å RMSD from the crystal structure with a dipeptide analogue. Indeed, the two crystal structures of AlbC in the apo state and in the dipeptide analogue state used for the initial structures for MD simulations are very similar, with the RMSD being 0.9 Å computed for 169 Cα atoms.[20] The RMSD for AlbC in the intermediate state, with Phe1, is practically the same as for the apoprotein, demonstrating that the bound Phe1 does not perturb the protein conformation. However, the RMSD computed for AlbC in the tRNA:AlbC complex is 1.33 Å, which is greater than 1.05 Å, demonstrating that for tRNA interactions the protein adopts a slightly different conformation. In particular, flexible loops between α2 and β3 (CL1) and between α6 and α7 are mainly contributing to the RMSD. However, the position of the backbone of the β-strands and α-helices, including α4, is very similar. The α2−β3 loop, so-called CL1,[20] contributes to the structure of the catalytic site, and in all available crystal structures in the tRNA free form, the conformation of this loop is very similar, suggesting that its flexibility may be involved in tRNA binding.

DISCUSSION

In this work, using a synergy of *in vivo* experiments and simulations, the AlbC:Phe-tRNA^{Phe} complex was investigated. Eleven initial models were obtained by rigid docking and through homology modeling based on the sequence and structure similarity with TyrRS, without imposing strict restraints on possible protein:tRNA interactions. The models were classified into four groups of possible tRNA binding positions. In all studied models, the interaction of the aminoacyl moiety with AlbC persisted in preliminary MD simulations, indicating that AlbC:tRNA binding relies on the interaction of the aminoacyl moiety with the active site residues as suggested previously.[12] Furthermore, the models from the two groups with the lowest protein:tRNA binding free energies converged to very similar structures of the AlbC:tRNA complex in MD simulations, which suggests that this binding pose is the solution to the binding problem. In all models with the lowest energy, AlbC interacts with the tRNA via helix α4 in agreement with the biochemical experiments. Notably, the analysis of the charge conservation showed that helix α4 has the total charge well conserved in the entire NYH subfamily. The average charge of helix α4 is +5.5 *e* (SD of 1.5 *e*), suggesting that the binding mode of tRNA is shared by the NYH CDPSs.

The model with the lowest binding free energy was then used to study the protein:tRNA complex in 1 μs long MD simulations. The model was stable as indicated by the RMSD and the protein:tRNA binding free energy. Analysis of individual residue contributions identified residues Lys46, Arg80, Arg87, Arg91, Lys94, Arg98, Arg99, and Arg102 as strongly contributing to the complex formation and stability. Asp95 was also identified by simulations to be implicated in tRNA interactions and further validated by *in vivo* experiments in this work. Arg80, Arg87, Arg91, Lys94, Arg98, Arg99, and Arg102 have been previously studied by mutations, demonstrating that all, except Arg87, are necessary for AlbC function.[12,15] All of these residues except Arg80 belong to helix α4. In MD simulations, helix α4 makes multiple ionic interactions by positively charged residues with the phosphate groups of both strands delimiting the major groove of the tRNA acceptor stem. Similar interactions were observed in other families of enzymes using aa-tRNA to form peptide bonds. Both FemX aminoacyl-transferases and aa-tRNA protein transferases recognize the cognate aa-tRNA via its acceptor stem.[23,24] Moreover, a proposed model for L/F transferase suggests that tRNA recognition occurs from a positive cluster located on a small solvent-exposed α-helix.[24,61]

However, interaction of FemX aminoacyl-transferases and aa-tRNA protein transferases with tRNA strongly depends on the tRNA sequence of the acceptor stem.[23,24] Interestingly, it was suggested that AlbC:tRNA interaction was different for the first or second tRNA.[12] For the binding of the first substrate, the interaction with the basic patch of helix α4 is essential. The binding of the second substrate is highly dependent on both the aminoacyl moiety and the tRNA sequence itself. It would involve the α6−α7 loop delimiting the P2 pocket but not helix α4.[12] AlbC distinguishes between the G^1-C^{72} and C^1-G^{72} pairs, similar to TyrRSs and FemX aminoacyl-transferases.[62−64] This indicates that the modeled AlbC:Phe-tRNA^{Phe} interaction corresponds to the pose of the first substrate as in our model, and there is no specific interaction with the G^1-C^{72} base pair, in agreement with the experimental evidence. However, in the 1 μs MD simulation, the phenylalanine of the bound tRNA inside the catalytic site predominantly occupied one of two positions compatible with Phe1 and Phe2 observed in the crystal structure (PDB entry 4Q24[20]) in the intermediate state. The conformational transitions between the Phe1 and Phe2 positions were accomplished without significant conformational rearrangements of the protein atoms. In MD simulations, the phenylalanyl group of tRNA spends around 2 times more time in the Phe1 position than in the Phe2 position, in agreement with the previous suggestion that Phe1 is more specific for the phenylalanine than Phe2, because the second substrate type is less strict in AlbC. This suggests that with the same tRNA binding mode, the phenylalanine group in principle can be poised for the second step of the catalytic reaction, and both tRNAs for the first and second steps of the enzymatic reaction bind in a mode similar to that of AlbC. This conjecture will be tested in future studies.

Interestingly, the previous experiments demonstrate a significant decrease in the enzymatic activity or a complete

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

loss for *Nbra*-CDPS or AlbC, respectively, when the XYP and NYH motifs in these CDPSs were converted to the NYH and XYP motifs, respectively.[14] This clearly demonstrates that the difference in the XYP and NYH motifs is just a reflection of more profound differences between the two subfamilies. We could propose that this distinction may be contributed by the difference in interactions of CDPSs with tRNA. Indeed, the overall positive charge of helix α4, proposed in this work to be important for the interactions of NYH CDPS with tRNA, is not observed in the XYP family, suggesting that helix α4 is not significantly involved in the tRNA recognition in agreement with recent crystallographic and biochemical studies.[14]

Overall, the simulation results are corroborated by biochemical experiments. In particular, the AlbC variants with basic residues belonging to helix α4 to be important for tRNA binding identified through mutation and conservation in sequence alignment are also the key residues for the AlbC:Phe-tRNA^{Phe} complex in our simulations. Overall, the mechanism of recognition of tRNA by CDPS AlbC discovered in this work is expected to be pertinent for other members of the CDPS NYH subfamily, because the CDPS:tRNA interaction involves the CDPS secondary structure with a conserved positive charge in the subfamily. This is also in line with the fact that CDPS:tRNA binding involves the tRNA acceptor stem, which is observed for other noncanonical enzymes such as FemX aminoacyl-transferases and aa-tRNA protein transferases.[23,27]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.0c00761.

Details of homology modeling, Poisson−Boltzmann (PB) binding free energies for ZDOCK docking structures, force field parametrization, results for force field parametrization, Tables S1−S9, Figures S1−S9, and force field parameters (PDF)

Accession Codes

UniProtKB Q8GED7.

■ AUTHOR INFORMATION

Corresponding Author

Alexey Aleksandrov − *Laboratoire d'Optique et Biosciences (CNRS UMR7645, INSERM U1182), Ecole Polytechnique, Institut polytechnique de Paris, F-91128 Palaiseau, France;* orcid.org/0000-0002-8150-3931; Email: Alexey.Aleksandrov@polytechnique.edu

Authors

Anastasia Croitoru − *Laboratoire d'Optique et Biosciences (CNRS UMR7645, INSERM U1182), Ecole Polytechnique, Institut polytechnique de Paris, F-91128 Palaiseau, France*

Morgan Babin − *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette, France*

Hannu Myllykallio − *Laboratoire d'Optique et Biosciences (CNRS UMR7645, INSERM U1182), Ecole Polytechnique, Institut polytechnique de Paris, F-91128 Palaiseau, France*

Muriel Gondry − *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette, France;* orcid.org/0000-0003-0398-3404

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.0c00761

Funding

This work was supported by French National Research Agency Grant ANR-18-CE44-0002 to A.A.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

CINES (Grant 2018-A0040710436) is acknowledged for the generous allocation of computer time. The authors thank Alexandre Couëtoux for performing mutagenesis experiments and Nicolas Canu for his help with LC-MS experiments.

■ REFERENCES

(1) Gondry, M., Sauguet, L., Belin, P., Thai, R., Amouroux, R., Tellier, C., Tuphile, K., Jacquet, M., Braud, S., Courçon, M., Masson, C., Dubois, S., Lautru, S., Lecoq, A., Hashimoto, S., Genet, R., and Pernodet, J.-L. (2009) Cyclodipeptide synthases are a family of tRNA-dependent peptide bond−forming enzymes. *Nat. Chem. Biol. 5*, 414−420.

(2) Moutiez, M., Belin, P., and Gondry, M. (2017) Aminoacyl-tRNA-Utilizing Enzymes in Natural Product Biosynthesis. *Chem. Rev. 117*, 5578−5618.

(3) Bellezza, I., Peirce, M. J., and Minelli, A. (2014) Cyclic dipeptides: from bugs to brain. *Trends Mol. Med. 20*, 551−558.

(4) Bolognesi, M. L., Ai Tran, H. N., Staderini, M., Monaco, A., López-Cobeñas, A., Bongarzone, S., Biarnés, X., López-Alvarado, P., Cabezas, N., Caramelli, M., Carloni, P., Menéndez, J. C., and Legname, G. (2010) Discovery of a Class of Diketopiperazines as Antiprion Compounds. *ChemMedChem 5*, 1324−1334.

(5) Furukawa, T., Akutagawa, T., Funatani, H., Uchida, T., Hotta, Y., Niwa, M., and Takaya, Y. (2012) Cyclic dipeptides exhibit potency for scavenging radicals. *Bioorg. Med. Chem. 20*, 2002−2009.

(6) Kumar, N., Gorantla, J. N., Mohandas, C., Nambisan, B., and Lankalapalli, R. S. (2013) Isolation and antifungal properties of cyclo(D-Tyr-L-Leu) diketopiperazine isolated from Bacillus sp. associated with rhabditid entomopathogenic nematode. *Nat. Prod. Res. 27*, 2168−2172.

(7) Kwak, M.-K., Liu, R., Kwon, J.-O., Kim, M.-K., Kim, A. H., and Kang, S.-O. (2013) Cyclic dipeptides from lactic acid bacteria inhibit proliferation of the influenza A virus. *J. Microbiol. 51*, 836−843.

(8) Nicholson, B., Lloyd, G. K., Miller, B. R., Palladino, M. A., Kiso, Y., Hayashi, Y., and Neuteboom, S. T. C. (2006) NPI-2358 is a tubulin-depolymerizing agent: in-vitro evidence for activity as a tumor vascular-disrupting agent. *Anti-Cancer Drugs 17*, 25−31.

(9) Mishra, A., Choi, J., Choi, S.-J., and Baek, K.-H. (2017) Cyclodipeptides: An Overview of Their Biosynthesis and Biological Activity. *Molecules 22*, 1796.

(10) Rhee, K.-H. (2004) Cyclic dipeptides exhibit synergistic, broad spectrum antimicrobial effects and have anti-mutagenic properties. *Int. J. Antimicrob. Agents 24*, 423−427.

(11) Lautru, S., Gondry, M., Genet, R., and Pernodet, J.-L. (2002) The Albonoursin Gene Cluster of S. noursei. *Chem. Biol. 9*, 1355−1364.

(12) Moutiez, M., Seguin, J., Fonvielle, M., Belin, P., Jacques, I. B., Favry, E., Arthur, M., and Gondry, M. (2014) Specificity determinants for the two tRNA substrates of the cyclodipeptide synthase AlbC from Streptomyces noursei. *Nucleic Acids Res. 42*, 7247−7258.

(13) Bonnefond, L., Arai, T., Sakaguchi, Y., Suzuki, T., Ishitani, R., and Nureki, O. (2011) Structural basis for nonribosomal peptide synthesis by an aminoacyl-tRNA synthetase paralog. *Proc. Natl. Acad. Sci. U. S. A. 108*, 3912−3917.

(14) Bourgeois, G., Seguin, J., Babin, M., Belin, P., Moutiez, M., Mechulam, Y., Gondry, M., and Schmitt, E. (2018) Structural basis for partition of the cyclodipeptide synthases into two subfamilies. *J. Struct. Biol. 203*, 17−26.

(15) Sauguet, L., Moutiez, M., Li, Y., Belin, P., Seguin, J., Le Du, M.-H., Thai, R., Masson, C., Fonvielle, M., Pernodet, J.-L., Charbonnier, J.-B., and Gondry, M. (2011) Cyclodipeptide synthases, a family of class-I aminoacyl-tRNA synthetase-like enzymes involved in non-ribosomal peptide synthesis. *Nucleic Acids Res. 39*, 4475−4489.
(16) Vetting, M. W., Hegde, S. S., and Blanchard, J. S. (2010) The structure and mechanism of the Mycobacterium tuberculosis cyclodityrosine synthetase. *Nat. Chem. Biol. 6*, 797−799.
(17) Jacques, I. B., Moutiez, M., Witwinowski, J., Darbon, E., Martel, C., Seguin, J., Favry, E., Thai, R., Lecoq, A., Dubois, S., Pernodet, J.-L., Gondry, M., and Belin, P. (2015) Analysis of 51 cyclodipeptide synthases reveals the basis for substrate specificity. *Nat. Chem. Biol. 11*, 721−727.
(18) Canu, N., Moutiez, M., Belin, P., and Gondry, M. (2020) Cyclodipeptide synthases: a promising biotechnological tool for the synthesis of diverse 2,5-diketopiperazines. *Nat. Prod. Rep. 37*, 312−321.
(19) Gondry, M., Jacques, I. B., Thai, R., Babin, M., Canu, N., Seguin, J., Belin, P., Pernodet, J.-L., and Moutiez, M. (2018) A Comprehensive Overview of the Cyclodipeptide Synthase Family Enriched with the Characterization of 32 New Enzymes. *Front. Microbiol. 9*, 46.
(20) Moutiez, M., Schmitt, E., Seguin, J., Thai, R., Favry, E., Belin, P., Mechulam, Y., and Gondry, M. (2014) Unravelling the mechanism of non-ribosomal peptide synthesis by cyclodipeptide synthases. *Nat. Commun. 5*, 5141.
(21) Schmitt, E., Bourgeois, G., Gondry, M., and Aleksandrov, A. (2018) Cyclization Reaction Catalyzed by Cyclodipeptide Synthases Relies on a Conserved Tyrosine Residue. *Sci. Rep. 8*, 7031.
(22) Aravind, L., de Souza, R. F., and Iyer, L. M. (2010) Predicted class-I aminoacyl tRNA synthetase-like proteins in non-ribosomal peptide synthesis. *Biol. Direct 5*, 48.
(23) Fonvielle, M., Chemama, M., Villet, R., Lecerf, M., Bouhss, A., Valery, J.-M., Etheve-Quelquejeu, M., and Arthur, M. (2009) Aminoacyl-tRNA recognition by the FemXWv transferase for bacterial cell wall synthesis. *Nucleic Acids Res. 37*, 1589−1601.
(24) Fung, A. W. S., Leung, C. C. Y., and Fahlman, R. P. (2014) The determination of tRNA ^{Leu} recognition nucleotides for *Escherichia coli* L/F transferase. *RNA 20*, 1210−1222.
(25) Bouhss, A., Josseaume, N., Allanic, D., Crouvoisier, M., Gutmann, L., Mainardi, J.-L., Mengin-Lecreulx, D., van Heijenoort, J., and Arthur, M. (2001) Identification of the UDP-MurNAc-Pentapeptide:l-Alanine Ligase for Synthesis of Branched Peptidoglycan Precursors in Enterococcus faecalis. *J. Bacteriol. 183*, 5122−5127.
(26) Hegde, S. S., and Shrader, T. E. (2001) FemABX Family Members Are Novel Nonribosomal Peptidyltransferases and Important Pathogen-specific Drug Targets. *J. Biol. Chem. 276*, 6998−7003.
(27) Watanabe, K., Toh, Y., Suto, K., Shimizu, Y., Oka, N., Wada, T., and Tomita, K. (2007) Protein-based peptide-bond formation by aminoacyl-tRNA protein transferase. *Nature 449*, 867−871.
(28) Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol. 7*, 539.
(29) Schrödinger, LLC. *The PyMOL Molecular Graphics System*, version 2.0.
(30) Byrne, R. T., Jenkins, H. T., Peters, D. T., Whelan, F., Stowell, J., Aziz, N., Kasatsky, P., Rodnina, M. V., Koonin, E. V., Konevega, A. L., and Antson, A. A. (2015) Major reorientation of tRNA substrates defines specificity of dihydrouridine synthases. *Proc. Natl. Acad. Sci. U. S. A. 112*, 6033−6037.
(31) Harrington, K. M., Nazarenko, I. A., Dix, D. B., Thompson, R. C., and Uhlenbeck, O. C. (1993) In vitro analysis of translational rate and accuracy with an unmodified tRNA. *Biochemistry 32*, 7617−7622.
(32) Mikkelsen, N. E., Johansson, K., Virtanen, A., and Kirsebom, L. A. (2001) Aminoglycoside binding displaces a divalent metal ion in a tRNA−neomycin B complex. *Nat. Struct. Biol. 8*, 510−514.
(33) Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and MacKerell, A. D. (2012) Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput. 8*, 3257−3273.
(34) Huang, J., and MacKerell, A. D. (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem. 34*, 2135−2145.
(35) Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys. 79*, 926−935.
(36) Neria, E., Fischer, S., and Karplus, M. (1996) Simulation of activation free energies in molecular systems. *J. Chem. Phys. 105*, 1902−1921.
(37) 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., and Karplus, M. (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B 102*, 3586−3616.
(38) Pierce, B. G., Wiehe, K., Hwang, H., Kim, B.-H., Vreven, T., and Weng, Z. (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics 30*, 1771−3.
(39) Iwakiri, J., Hamada, M., Asai, K., and Kameda, T. (2016) Improved Accuracy in RNA-Protein Rigid Body Docking by Incorporating Force Field for Molecular Dynamics Simulation into the Scoring Function. *J. Chem. Theory Comput. 12*, 4688−4697.
(40) Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem. 26*, 1781−802.
(41) Darden, T. (2001) Treatment of Long-Range Forces and Potential. In *Computational Biochemistry & Biophysics* (Becker, O. M., MacKerell, A. D., Jr., Roux, B., and Watanabe, M., Eds.) Marcel Dekker, New York.
(42) Hart, K., Foloppe, N., Baker, C. M., Denning, E. J., Nilsson, L., and Mackerell, A. D., Jr. (2012) Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J. Chem. Theory Comput. 8*, 348−362.
(43) Huang, J., and MacKerell, A. D. (2013) CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem. 34*, 2135−45.
(44) Byrne, R. T., Konevega, A. L., Rodnina, M. V., and Antson, A. A. (2010) The crystal structure of unmodified tRNA Phe from Escherichia coli. *Nucleic Acids Res. 38*, 4154−4162.
(45) 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., and Karplus, M. (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem. 30*, 1545−1614.
(46) Gilson, M. K., Sharp, K. A., and Honig, B. H. (1988) Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem. 9* (4), 327−335.
(47) Im, W., Beglov, D., and Roux, B. (1998) Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. *Comput. Phys. Commun. 111*, 59−75.
(48) Gilson, M. K., and Honig, B. H. (1987) Calculation of electrostatic potentials in an enzyme active site. *Nature 330*, 84−86.
(49) Kuhn, B., and Kollman, P. A. (2000) Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

Mechanics and Continuum Solvent Models. *J. Med. Chem. 43*, 3786−3791.
(50) Sham, Y. Y., Muegge, I., and Warshel, A. (1998) The Effect of Protein Relaxation on Charge-Charge Interactions and Dielectric Constants of Proteins. *Biophys. J. 74*, 1744−1753.
(51) Schutz, C. N., and Warshel, A. (2001) What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Struct., Funct., Genet. 44*, 400−417.
(52) Genheden, S., Kuhn, O., Mikulskis, P., Hoffmann, D., and Ryde, U. (2012) The Normal-Mode Entropy in the MM/GBSA Method: Effect of System Truncation, Buffer Region, and Dielectric Constant. *J. Chem. Inf. Model. 52*, 2079−2088.
(53) Ma, J. (2005) Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. *Structure 13*, 373−380.
(54) Cummins, P. L., Ramnarayan, K., Singh, U. C., and Gready, J. E. (1991) Molecular dynamics/free energy perturbation study on the relative affinities of the binding of reduced and oxidized NADP to dihydrofolate reductase. *J. Am. Chem. Soc. 113*, 8247−8256.
(55) Aleksandrov, A., Schuldt, L., Hinrichs, W., and Simonson, T. (2008) Tet Repressor Induction by Tetracycline: A Molecular Dynamics, Continuum Electrostatics, and Crystallographic Study. *J. Mol. Biol. 378*, 898−912.
(56) Simonson, T., Carlsson, J., and Case, D. A. (2004) Proton binding to proteins: pK(a) calculations with explicit and implicit solvent models. *J. Am. Chem. Soc. 126*, 4167−4180.
(57) Canu, N., Belin, P., Thai, R., Correia, I., Lequin, O., Seguin, J., Moutiez, M., and Gondry, M. (2018) Incorporation of Non-canonical Amino Acids into 2,5-Diketopiperazines by Cyclodipeptide Synthases. *Angew. Chem., Int. Ed. 57*, 3118−3122.
(58) Felder, C. E., Prilusky, J., Silman, I., and Sussman, J. L. (2007) A server and database for dipole moments of proteins. *Nucleic Acids Res. 35*, W512−W521.
(59) Aleksandrov, A., and Simonson, T. (2008) Molecular Dynamics Simulations of the 30S Ribosomal Subunit Reveal a Preferred Tetracycline Binding Site. *J. Am. Chem. Soc. 130*, 1114−1115.
(60) Yaremchuk, A. (2002) Class I tyrosyl-tRNA synthetase has a class II mode of cognate tRNA recognition. *EMBO J. 21*, 3829−3840.
(61) Suto, K., Shimizu, Y., Watanabe, K., Ueda, T., Fukai, S., Nureki, O., and Tomita, K. (2006) Crystal structures of leucyl/phenylalanyl-tRNA-protein transferase and its complex with an aminoacyl-tRNA analog. *EMBO J. 25*, 5942−5950.
(62) Bedouelle, H. (2013) *Tyrosyl-tRNA Synthetases*, Landes Bioscience. https://www.ncbi.nlm.nih.gov/books/NBK6553/ (accessed 2020-06-08).
(63) Bonnefond, L., Giegé, R., and Rudinger-Thirion, J. (2005) Evolution of the tRNATyr/TyrRS aminoacylation systems. *Biochimie 87*, 873−883.
(64) Kobayashi, T., Nureki, O., Ishitani, R., Yaremchuk, A., Tukalo, M., Cusack, S., Sakamoto, K., and Yokoyama, S. (2003) Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. *Nat. Struct. Mol. Biol. 10*, 425−432.

https://dx.doi.org/10.1021/acs.biochem.0c00761
Biochemistry XXXX, XXX, XXX−XXX

CONCLUSION

In this work, using molecular modelling tools, I studied the interaction of cyclodipeptide synthases with their first substrate, the aminoacylated tRNA. Initial interaction poses were obtained with rigid docking and further refined with MD simulations to propose a model of interaction for the complex. CHARMM force field development was needed for the parameters of the phenylalanyl group covalently bonded to tRNA. In collaboration with an experimental group, we showed that the proposed model is compatible with the experimental data.

This study is a fine example of the necessity of force field development in order to accommodate the chemical modifications of standard biopolymers. The next step is to improve the force field optimization method presented in **Chapter 5**.

Chapter 5
DEVELOPMENT OF A NEW METHOD FOR BOND AND VALENCE ANGLE BONDED TERMS PARAMETRIZATION

One of the important features of the force field is the transferability of parameters which implies that the same parameters can be used to model an ensemble of similar molecules, rather than creating an individual set of parameters for each individual molecule. Generally, parametrization of force fields is performed in small molecules and then adapted to the larger complete molecule. Force fields for small molecules are constructed in incremental parametrization procedures, where parameters developed previously are retained for novel molecules, followed by optimization of missing, not previously optimized parameters. However, equilibrium QM and MM geometries of molecules can deviate due to parameters transferred from existing molecules in the force field. We demonstrate that conventional parametrization methods based on fitting QM energies and/or forces to derive parameters for bond and angle terms produce largely suboptimal force constants when MM and QM equilibrium structures deviate even slightly. We propose a new method to derive force field parameters based on the PES scans where a structural deviation between QM and MM optimized geometries is explicitly allowed during parametrization.

The test of the developed method was performed on a diverse set of 32 small molecules. Starting from random initial force constant, the parameters for bond and angle terms optimized by the new method converge to the optimal value. We further demonstrate using the test molecules that the bond and angle parameters produced by the new method are largely transferable, with the force constants optimized in different molecules deviating by less than 2% on average. An additional test was performed for normal modes. The new method also improves the agreement for the normal modes for all molecules in the set, reducing the average error in the reproduction of QM normal mode frequencies for optimized parameters compared to initial ones. The new method will allow parametrization of molecules under structural deviations, common for force fields for small molecules, producing robust

and transferable parameters. The presented chapter is a based on a submitted article.

In this work, we mainly focus on bond and angle terms of the bonded part of the total FF energy. In widely used Class I additive force fields these terms are modeled by harmonic functions to describe deformations along bonds and angles around their equilibrium values. Parameters for these terms can be obtained by reproducing the experimental or/and QM vibrational spectrum, Hessian matrix[138,139], and deformation energies and forces.[140] In the later methods, also known as the Force Matching methods, MM parameters are fitted to reproduce the forces in non-equilibrium structures, which can be generated, for example, by classical MD simulations.[140–144] For the CHARMM force field, historically, to parametrize stiff degrees of freedom a symbolic potential energy distribution (PED) analysis was performed in the internal coordinate space.[145] This allows estimating relative contributions of the valence coordinates to frequencies. These contributions are computed using the MM Hessian calculated using the trial parameters, and compared with the corresponding QM PED; parameters are iteratively varied until satisfactory agreement is reached. In practice, the fitting is difficult since QM and MM frequencies of a normal mode as well as QM and MM contributions of internal coordinates to the same normal mode are different. With this, the quality of the fit is difficult to quantify and in addition, one needs to define a non-unique mapping between internal coordinates and normal modes, which is difficult to automate.

Apart from PED analysis, a method to determine force constants by three-point PES scans was used for CGenFF when the assignment of the internal coordinate contributions to the vibrations was ambiguous.[113] This method is also implemented in the Force Field Toolkit (ffTk), a VMD plugin[146] that can be used to parametrize the CHARMM force field for small molecules.[147] In this method, a small distortion in two opposing directions is generated and the corresponding increase in potential energy relative to the undistorted conformation is computed. The QM Hessian is used to compute QM energy for the small distortions about the minimized geometry. The energies are scaled to improve the agreement with experimental vibrational frequencies.[113] Since no optimization is done for the deformed structures, in principle different sets of parameters can reproduce QM energies equally well. Thus, the parameter optimization problem is ill-defined in this case, and requires a restraining strategy. Different such restraint strategies have been proposed.[113,147–149] However, introducing such artificial restraints may result in a poor transferability of

parameters for molecules that were not used for the optimization. In particular, bond and angle parameters are developed typically only in one molecule and used for all other molecules in the chemical universe sharing the same term defined by atom types.

Equilibrium QM and MM geometries of molecules can deviate due to parameters transferred from existing molecules in the force field, even with FF parameters for new terms optimized against QM reference data. For example, in the previous work on the parametrization of the large set of nonstandard amino acids we found that bonds and angles deviate on average by 0.02 Å and 2°, respectively, for a large set of 189 compounds after optimization of new parameters not existing in CGenFF. In this work, we demonstrate that, while these MM structural inconsistences relative to QM optimized structures can be negligible for applications, they strongly impact the quality of new parameters optimized in novel molecules. We further show that the conventional methods to derive parameters for bond and angle terms produce largely suboptimal force constants when MM and QM equilibrium structures deviate even slightly. This problem arises if the same structures (for example, QM optimized structures) are used during parametrization for QM and MM calculations, or QM and MM structures have the same value of the deformed bond or angle. We further developed and tested a new method to derive force field parameters based on the PES scans where a structural deviation between QM and MM structures is explicitly allowed and show that the new method produces stable and transferable parameters for bond and angle terms without any need for additional restraints.

OPTIMIZATION AND COST FUNCTION

Typically, to optimize bonded parameters a cost function is constructed using energy differences between QM and MM structures, for example from PES scans, which is further minimized to give an optimal set of bonded parameters. In the simplest form, the RMS deviation between QM and MM energies for a set of structures can be used for the cost function:

$$F^{ener} = \sqrt{\frac{1}{N}\sum_i (E^{QM}(\boldsymbol{x}_i) - E^{MM}(\boldsymbol{x}_i))^2}, \text{ [Eq 5.4]}$$

where the sum is over a set of N structures; E^{QM} and E^{MM} are QM and MM energies, respectively, computed using the same set of coordinates, $\boldsymbol{x}_i$. The structures can be generated in different ways, however, in this work, we will focus on potential energy surface scans, described above. For calculations with the force field model,

structures can correspond to QM optimized structures, or typically, they can be re-optimized with the force field model. Thus, QM and MM structures can be different, however, the scanned valence coordinate (for example a dihedral angle) has the same value in the QM and MM structures. In this case, the RMS deviation is given by:

$$F^{ener} = \sqrt{\frac{1}{N}\sum_i (E^{QM}(\boldsymbol{x}_i^{QM}) - E^{MM}(\boldsymbol{x}_i^{MM}))^2}, \text{ [Eq 5.5]}$$

where the two sets of coordinates, $\boldsymbol{x}_i^{QM}$ and $\boldsymbol{x}_i^{MM}$ have the same value along the scanned internal degree of freedom, k: $x_{i,k}^{QM} = x_{i,k}^{MM}$.

NORMAL MODE ANALYSIS

The quality of the optimized parameters was tested using normal mode analysis (NMA). QM NMA was performed for the molecules in the data set using the optimized structures at the MP2/6-31G* level of theory. The structures were fully optimized with quadratically convergent SCF procedure[150] and there were no normal modes with negative frequencies. The correspondence between QM and MM normalized normal modes was determined based on the dot product before further comparison. For this purpose, normal modes were considered as collinear if their absolute dot product is >0.5. If a QM normal mode is contributed by several CHARMM normal modes, the MM mode with the largest dot product was considered.

To characterize the fit between QM and MM frequencies, Mean Percentage Absolute Error (MPAE) was calculated using:

$$\text{MPAE}= \frac{100\%}{n}\sum_i^n \left| \frac{\alpha\nu_i^{QM} - \nu_j^{MM}}{\alpha\nu_i^{QM}} \right|, \text{ [Eq 5.6]}$$

where n is number of colinear NM; ν_i^{QM} and ν_j^{MM} are the frequency of QM NM and the frequency of corresponding collinear CHARMM and QM NM; $\alpha = 0.9432$ is the vibrational scaling factor.[151] MPAE was adapted from a previous work.[152] The bonded parameters were optimized with the scaling factor of α^2 applied to QM energies to obtain scaling factor of α for QM frequencies. The comparison was also done with the QM frequencies scaled by this factor in Equation 11. Bonded parameters were derived starting from initial CHARMM parameters for force constants and equilibrium values obtained from *ab initio* optimized geometry of the molecule.

TEST MOLECULES AND ATOM TYPES

To test different parametrization methods, 32 molecules were selected with available CHARMM parameters. The set of molecules comprises molecules with diverse chemical structures and includes three, four, five and six-atom ring structures. In total, the set contains 74 unique bond terms and 127 unique angle terms with an average of 4 and 6 terms per molecule respectively. Chemical structures of these molecules are shown in Figure 5.1 and their chemical names and formulae are given in Table 5.1.

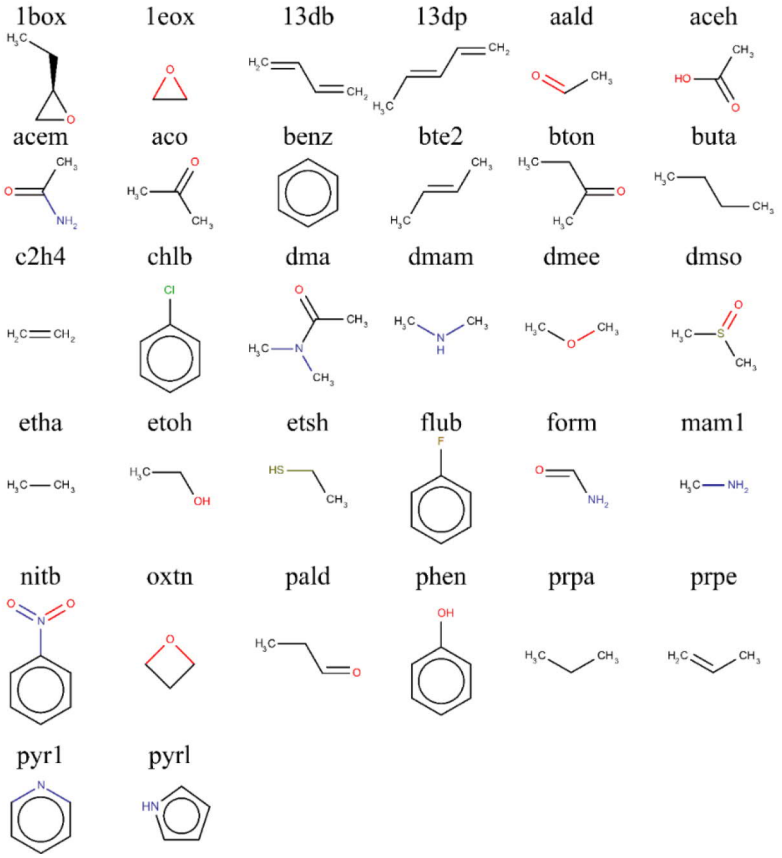


Figure 5.1. Chemical structures of 32 molecules used in this work. 2D representations were prepared with MarvinSketch.[129]

Table 5.1 Nomenclature of the 32 molecules test set.

CHARMM residue name	Chemical formula	IUPAC name
13db	C4H6	1,3-dibutene
13dp	C5H8	1,3-dipentene
1box	C4H8O	1-butene oxide
1eox	C2H4O	1-ethylene oxide
aald	C2H4O	acetaldehyde
aceh	C2H4O2	acetic acid
acem	C2H5NO	acetamide
aco	C3H6O	acetone
benz	C6H6	benzene
bte2	C4H8	2-butene
bton	C4H8O	butanone
buta	C4H10	butane
c2h4	C2H4	ethylene
chlb	C6H5Cl	chlorobenzene
dma	C4H9NO	dimethylacetamide
dmam	C2H7N	dimethylamine
dmee	C2H6O	dimethylether
dmso	C2H6OS	dimethylsulfoxide
etha	C2H6	ethane
etoh	C2H6O	ethanol
etsh	C2H6S	ethanethiol
flub	C6H5F	fluorobenzene
form	CH3NO	formamide
mam1	CH5N	methylamine
nitb	C6H5NO2	nitrobenzene
oxtn	C3H6O	oxetane
pald	C3H6O	propionaldehyde
phen	C6H6O	phenol
prpa	C3H8	propane
prpe	C3H6	propene
pyr1	C5H5N	pyridine
pyrl	C4H5N	pyrrole

The force field parameters for the molecules in the test set were taken from CGenFF.[95] For simplicity, these parameters will be further referred as initial CHARMM parameters. The geometry of molecules was generated from the existing tables of internal coordinates in the CGenFF force field files. The geometries were further optimized at the MP2/6-31G* model chemistry, or MP2/6-311G(d) model chemistry for anionic molecules. Adiabatic potential energy surface (PES) scans with

QM were performed along selected degrees of freedom as described above. MM calculations were performed with CHARMM program and QM calculation were performed with Gaussian09.[106]

RESULTS

Deviation between QM and MM relaxed structures leads to suboptimal force constants

In this section, the arguments will be illustrated on selected molecules. Figure 5.2A shows the PE surface for C-C bond stretching in ethylene computed with QM and optimal CHARMM parameters. The Root Mean Square (RMS) deviation between QM and MM energies computed with Formula 5 is shown in Figure 5.2 (panels B and C) computed with different force field equilibrium bond distance values, b_0; and with different force constant values, K_b. For these calculations, the optimal values for the term were used for b_0 and K_b, respectively; the other terms optimized in PES scans were also treated with the optimal parameters. The MM structures were optimized using the same value for the valence coordinate (bond or angle), which was used for the QM optimizations, i.e. $x_{i,k}^{QM} = x_{i,k}^{MM}$. As it can be seen, even relatively small deviations in b_0 lead to large deviations between QM and MM energies, and thus the RMS deviation. For example, a deviation of just 0.05 Å from the optimal value for b_0 leads to a value of 2.6 kcal·mol^{-1} for the RMS deviation (a deviation of 0.1 Å leads to ~8 kcal·mol^{-1}). In contrast, relatively large deviations in K_b produce only a small increase in the RMS deviation. For example, in ethylene, reducing the force constant by 300 kcal·mol^{-1}·Å^{-2} would increase the RMS deviation only by 0.35 kcal·mol^{-1}. Thus, if the RMS deviation is used for the cost function, the optimization would be largely balanced toward a better equilibrium distance b_0 to improve the RMS deviation, while the quality of the force constants could be sacrificed. In practice, this is strongly undesirable since force constants are important to reproduce the molecular flexibility. This can be further demonstrated on the heatmap of the RMS deviation between QM and MM energies computed with different b_0 and K_b on Figure 5.3A. With a deviation of just 0.02 Å in the equilibrium value, the optimized value for the force constant is on order of 300 kcal·mol^{-1}, which is ~300 kcal·mol^{-1} off from the optimal force constant needed to reproduce the C-C bond stretching in ethylene.

We shall consider another example of the valence angle in acetaldehyde. Figure 5.2 panel A shows the PE surface for O-C-C valence angle bending in acetaldehyde computed with QM and optimal CHARMM parameters; while the RMS

deviation between QM energies and MM energies calculated at different equilibrium angle values, θ_0 and force constants, K_a is shown in panels B and C of the same figure. The Figure 5.3B shows the heatmap of the RMS deviation between QM and MM energies at different θ_0 and K_a. It can be seen that similar to the bond term relatively small deviations in θ_0 increase significantly the difference in QM and MM energies and hence the RMS deviation. For example, the RMS deviation of 0.6 kcal·mol^{-1} corresponds to just less than 3° deviation from the optimal equilibrium valence angle and the same RMSD corresponds to 38 kcal·mol^{-1} deviation for the force constant K_a (over 100%).

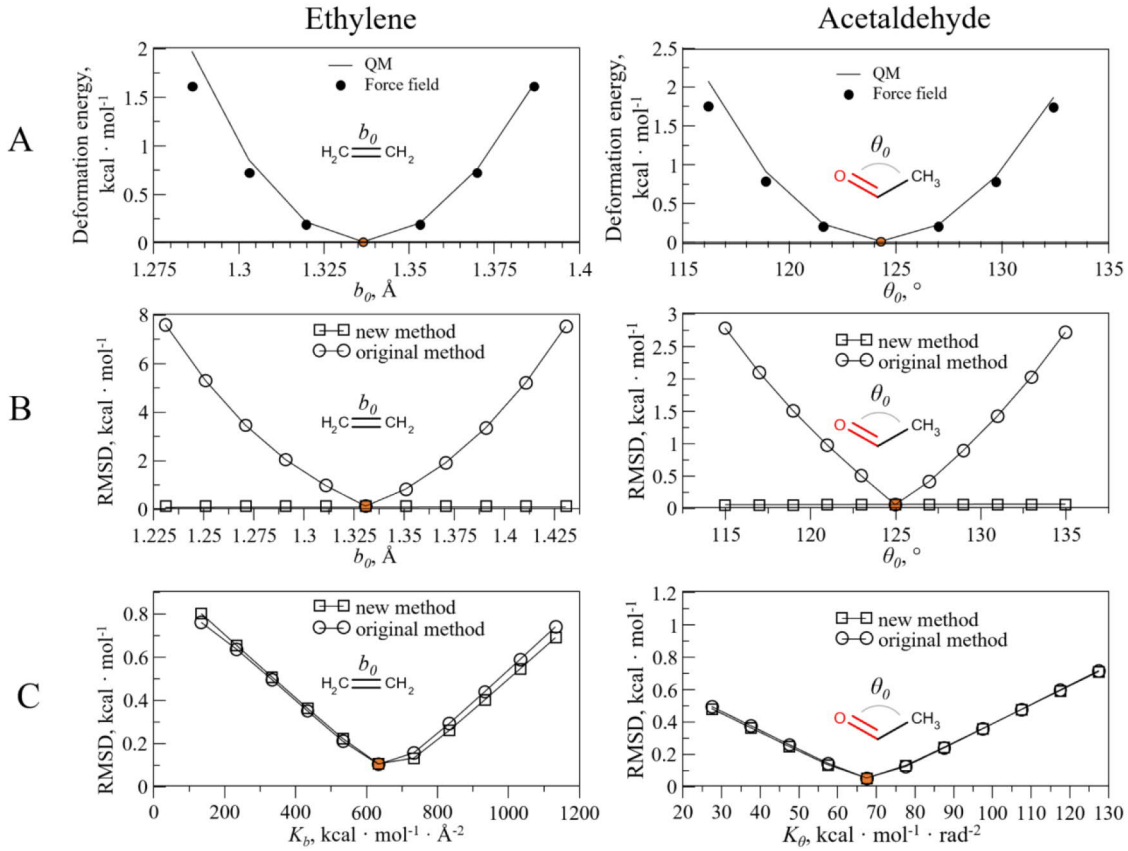


Figure 5.2. Potential energy surfaces and energy RMS deviations as a function of the force field parameters. (A) Left, the PE surface for the C-C bond stretching in ethylene, and right, for the O-C-C angle bending for acetaldehyde. (B) The RMS deviation between QM and CHARMM energies of structures from the PES scans as a function of the equilibrium bond distance, b_0, and valence angle, θ_0. (C) RMS deviation between QM and CHARMM energies as a function of the force constant for the bond term in ethylene (left) and the force constant of the angle term in acetaldehyde, right, respectively. The optimal values for the equilibrium bond distance, valence angle and force constants are colored in orange.

The above reasoning was illustrated PES energies; however, they can be generalized to force matching methods. The average RMS deviation between QM and MM atomic forces in structures from the PES scans is shown in Figure 5.3. Overall, the behavior of the force RMS deviation is very similar to the energy RMS deviation,

i.e. small deviations from the optimal equilibrium value for the bond in ethylene and the valence angle in acetaldehyde lead to significantly suboptimal force constants relative to the QM model. We find that deviations in equilibrium values for the bond and angle lead to the same force constants: with a deviation of 0.02 Å in the equilibrium value, the optimized value for the force constant is on order of 300 kcal·mol^{-1}, which is ~300 kcal·mol^{-1} off from the optimal force constant needed to reproduce the C-C bond stretching in ethylene. For the valence angle in acetaldehyde 3° deviation from the optimal equilibrium valence angle results in ~40 kcal·mol^{-1} deviation for the force constant K_a.

In both considered cases for the bond and angle terms, structural deviations between the QM and MM optimized structures along the corresponding bond and angle lead to force constants that are smaller compared to the values optimized in the absence of these deviations. In general, the resulting MM model under a structural inconsistency as demonstrated in Figure 5.3, is softer than the QM model, and in other words, it allows larger amplitudes of deformations at the same energy values compared to the QM model.

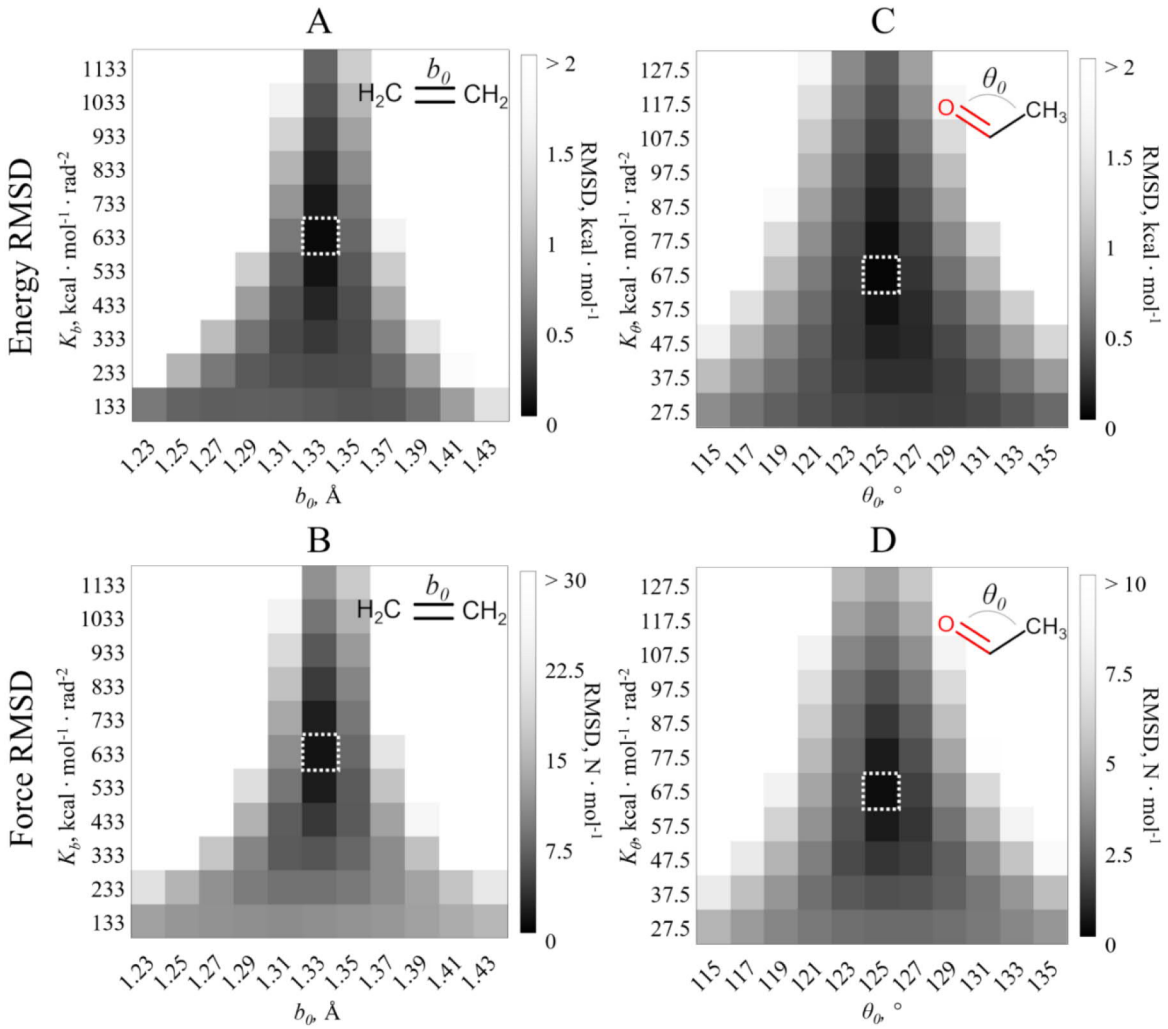


Figure 5.3. The RMS deviation between QM and CHARMM energies and forces of structures from PES scans as a function of FF parameters. (A) The energy RMS deviation and (B) the force RMS deviation as a function of the equilibrium bond distance, b_0, and force constant, K_b, of the C–C bond term in ethylene. (C) The energy RMS deviation and (D) the force RMS deviation as a function of the equilibrium angle, θ_0, and force constant, K_θ, of the O–C–C angle term in acetaldehyde. The optimal values for the force field parameters are shown in a white dotted square.

Modification of the cost function

To remove the strong dependence on equilibrium parameters, demonstrated above, we allow structural deviations along the scanned degree of freedom, where energy differences are computed between different structures used for QM and MM calculations. These QM and MM structures are now different, in principle, in all coordinates. One of the requirements for such MM structures is that QM and MM structures, which energies are compared in Formula 5 are close to each other in the configurational space, in another words the MM and QM PES approximately match. We use a similar method described above where the MM structures are optimized

with one internal coordinate constrained, however, in the new method the value x_k^{MM} of the constrained coordinate, k, is allowed to deviate from x_k^{QM} and is given by:

$$x_k^{MM} = x_{0,k}^{MM} + (x_k^{QM} - x_{0,k}^{QM}), \text{ [Eq 5.7]}$$

where $x_{0,k}^{MM}$ and $x_{0,k}^{QM}$ are values in the MM and QM structures optimized structures without any constraints; x_k^{QM} is the value in the corresponding QM structure. The terms in the later formula can be regrouped:

$$x_k^{MM} = x_k^{QM} + \Delta x_{0,k}^{QM-MM}, \text{ [Eq 5.8]}$$

where $\Delta x_{0,k}^{QM-MM} = x_{0,k}^{MM} - x_{0,k}^{QM}$ is the deviation in the optimal values in the QM and MM optimized structures, which can be considered as a correction term. As it will be illustrated in the next section, the RMS deviation computed with the new definition of x_k^{MM} depends on the MM equilibrium values only slightly. As mentioned before, one of requirements is that the QM and MM structures should be close on in the configurational space, which can be achieved by introducing an additional term, also needed to provide a bias in optimization of equilibrium parameters. In particular, we introduced an additional simple term given by:

$$F^{eq} = \sqrt{\frac{1}{N}\sum_i (x_{k,i}^{MM} - x_{k,i}^{QM})^2}, \text{ [Eq 5.9]}$$

where $x_{k,i}^{MM}$ and $x_{k,i}^{QM}$ is the scanned internal coordinate in the i MM and QM optimized structure, respectively.

Additional term for angle terms

Let's consider deformations along the valence angles involving the ortho hydrogen atom of pyridine shown in Figure 5.4. In a previous work the conclusion was drawn that defining force constants for the valence angles for this atom is an ill-defined problem.[113] To derive this conclusion the ring structure was assumed to be rigid with all reference angles having 120°. In this case, indeed, the in-plane hydrogen bending can be described by only one valence angle, and energy in Formula 2 can be expressed as one harmonic term with the force constant given by the sum of the force constants of the two angle terms involving the hydrogen atom ($K_{effective} = K_1 + K_2$). However, this is not the case if minimization is done as the pyridine ring is not rigid. The angles with the atomic center are defined as shown in Figure 5.4. In practice, with the angle θ_1 between N-C-H bent out of its equilibrium value, the other two angles θ_2 (C-C-N) and θ_3 (C-C-H) would also assume values different from their

values in the minimized structures. In particular, the valence angle θ_3 would be different than 120° with in-plane hydrogen bending. Figure 5.4A shows the dependence of θ_2 and θ_3 on θ_1. For example, with θ_1 bending of 5°, the structure is bent along θ_2 and θ_3 by around 1° and 4°, respectively.

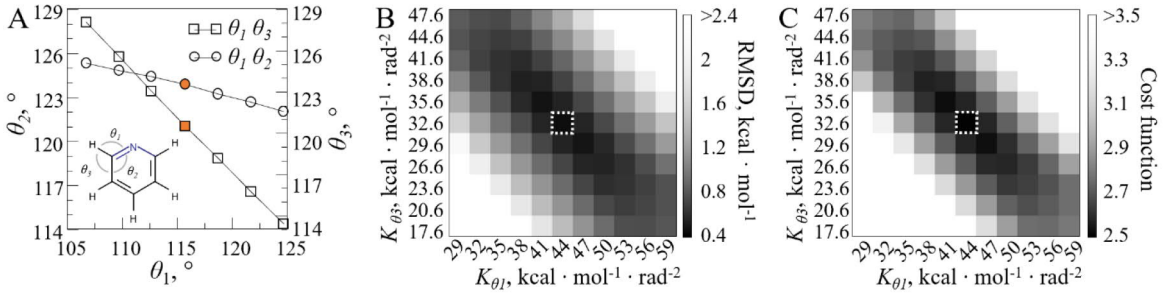


Figure 5.4. Angle bending involving ortho-hydrogen in pyridine. (A) Angles involving ortho-hydrogen in pyridine in the PES scan along θ_1 angle between N–C–H. Values corresponding to the minimum energy structure are in orange. (B) RMS deviation between QM and CHARMM energies from the PES scan as a function of the force constants of the two angle terms defined for ortho-hydrogen in pyridine (θ_1 and θ_3) (C) The cost function that includes the new term proposed in this work as a function of the two force constants K_1 and K_3. The optimal values for the force constants are marked by dotted line squares.

At a particular bending along θ_1 in the MM optimized structure, the other two angles depend on the force constants of the corresponding angle terms K_{θ_2} and K_{θ_3}, which should be sufficient to define these force constants. This conjecture we will test numerically in the next section. To improve further the distribution of force constants of angle terms, we tested an additional term to the cost function. We note that at a particular value of the angle, θ_1, the position of the atom N, and thus angles θ_2 and θ_3 depend on the corresponding force constants in the MM model. Thus, including the deformations along the adjacent angles, in principle, is expected to improve the distribution of the force constant. For an atomic center, which has angle terms to parametrize, the following deformation-based term is included to the cost function in addition to the restraints on the scanned degree of freedom given by Formula 8:

$$F^{def} = \sqrt{\sum_j \sum_i ((\theta_{j,i}^{MM} - \theta_{j,i+1}^{MM}) - (\theta_{j,i}^{QM} - \theta_{j,i+1}^{QM}))^2}, \text{ [Eq 5.10]}$$

where the summing is done over all structures and all angles involving this atomic center with adjustable parameters in the FF model. In this formula, angles in the MM and QM structures, are subtracted between subsequent structures on PES similar to Equation 6 to remove strong dependences on the equilibrium values. The total cost function is a sum of different contributions given by Formulae 5.5, 5.9 and 5.10:

$$F^{cost} = \omega_{ener} \cdot F^{ener} + \omega_{eq} \cdot F^{eq} + \omega_{def} \cdot F^{def}, \text{ [Eq 5.11]}$$

where ω_{ener} , ω_{def}, and ω_{eq}, are corresponding weights. These weights were defined based on our previous experience with parametrizing a large set of molecules. In particular, the weights were defined in such a way that the terms in Formula 10 give equal contributions (of a unity by definition) with the RMS deviations for energies and structural parameters obtained in the previous work.[102] The values for the weights are given in Table 5.2.

Table 5.2 Weights for the cost function

Term	Weight, ω
F^{ener}	10.0/3.3
$F^{eq}{}_{bond}$	66.7
$F^{eq}{}_{angle}$	0.67
F^{def}	0.87

Stability relative to the initial parameters

Since we employ optimization as a method to derive force field parameters, in principle, different sets of bonded parameters can be obtained starting from different initial values. To characterize the stability of optimized parameters with the new method we performed a numerical test where optimization was initiated from different initial force constants. Force constants, K_b for bond terms were assigned randomly from a wide range of values between 100 and 800 kcal·mol^{-1}·Å^{-2}; for angle terms, force constants, K_a were randomly assigned in the range of 10 to 100 kcal·mol^{-1}·rad^{-2}. The optimization was repeated five times starting from different random force constants; initial equilibrium bond lengths and angles were taken from QM structures. Relative and absolute standard deviation (SD) for bond and angle terms averaged over terms in individual molecules from the data set are shown in Figure 5.4 and SD averaged over all molecules in the data set, is given in Table 5.3.

Table 5.3. Standard deviation for bonded term parameters averaged over 32 molecules in the data set.

	Force constants, K_a or K_b		Equilibrium parameters, b_0 or θ_0	
Optimized terms	Average RSD	SD	Average RSD	SD
Bonds	0.3 (0.5)%	1.29 (1.94)	0.01 (0.01)%	0.00011 (0.00018)
Angles without the new term[a]	1.2 (2.8)%	0.69 (1.91)	0.06 (0.14)%	0.07 (0.17)
Angles with the new term[b]	0.7 (1.4)%	0.40 (0.88)	0.13 (0.31)%	0.15 (0.36)

[a,b]For angles, force field parametrization was done without and with the additional deformation-based term given by Equation 9 included to the cost function, respectively; the standard deviations for the computed values are shown in parenthesis. K_b and K_a are in kcal·mol^{-1}·Å^{-2} and kcal·mol^{-1}·rad^{-2}, respectively; the equilibrium parameters, b_0 or θ_0, are in Å and °, respectively.

The results demonstrate that the new method produces practically the same force constants starting from very different initial parameters. The relative SD for the force constants for bond terms, K_b averaged over all bond terms (74 total) in 32 molecules is just 0.3%. For angle terms, the relative SD for K_a averaged over a total of 127 angle terms is also very small, 1.2%. The later value can be further improved to 0.7% by introducing the deformation-based term given by Formula 9 to the cost function. Though, this improvement is small on average, for some angle terms it can present a significant improvement, for example for oxetane and 1,3-dibutene it improves the relative SD from 14% to 6%, and 7% to ~0%, respectively. The equilibrium parameters deviate only insignificantly in five optimization tests. The relative SD is 0.01% and ~0.1% for the equilibrium bond length and angle. Overall, bond and angle parameters derived by the new method are very stable regardless of the initial values used for the optimization.

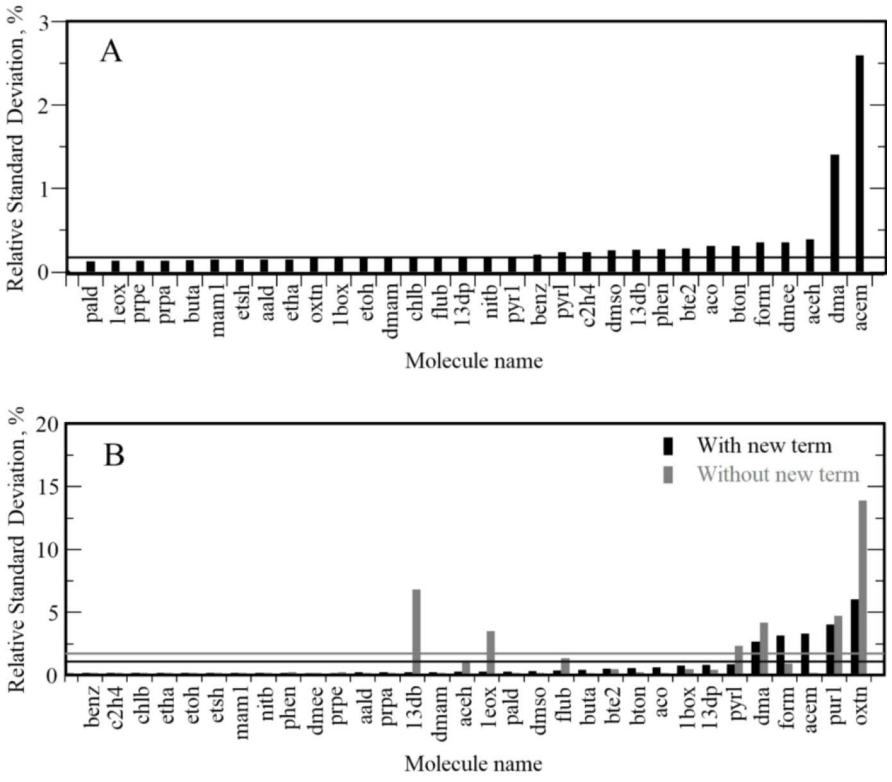


Figure 5.5. Robustness of optimized FF force constants relative to the initial parameters. The relative standard deviation of force constants for (A) bond and (B) angle terms is shown. Five optimizations started with random force constants were performed to compute the relative SD, which was averaged over bonded terms in individual molecules. The relative SD averaged over all molecules is shown as a horizontal line.

Transferability of optimized parameters

In the previous section we showed that the optimized parameters are robust and do not depend on initial parameters. In this section we will demonstrate that the optimized bonded parameters for bond and angle terms using the new method are transferable. For this, the comparison of the same bonded terms optimized in different molecules is performed. Indeed, 23 bond parameters and 24 angle parameters are shared by 2 to 17 molecules in the data set. For example, the bond term defined between atom types CG331-HGA3, i.e. between methyl hydrogen and carbon, is shared by 17 molecules, while the angle term, CG2R61- CG2R61-CG2R61, i.e. defined for the angle between carbon atoms in an aromatic ring is shared by 6 molecules. For this test, bond and angle terms were optimized using the new method; for angles, the optimization was done with and without the deformation-based term given by Equation 9. Bonds or angles in two or more molecules were identified and the SD and relative SD for the same bonded parameters optimized in different molecules were computed. The SD for bond and angle terms are shown in Figure 5.6

and given in Table 5.4 for individual terms. The following analysis does not include the angle term in two molecules having the three-atom ring (1box and 1eox), since by contrast to other molecules, a bending along the angle and stretching along the opposite bond in the three atom ring have two contributions: from the opposite bond and the contribution from the angle, and the bond term can be considered as an Urey-Bradley term for the angle.

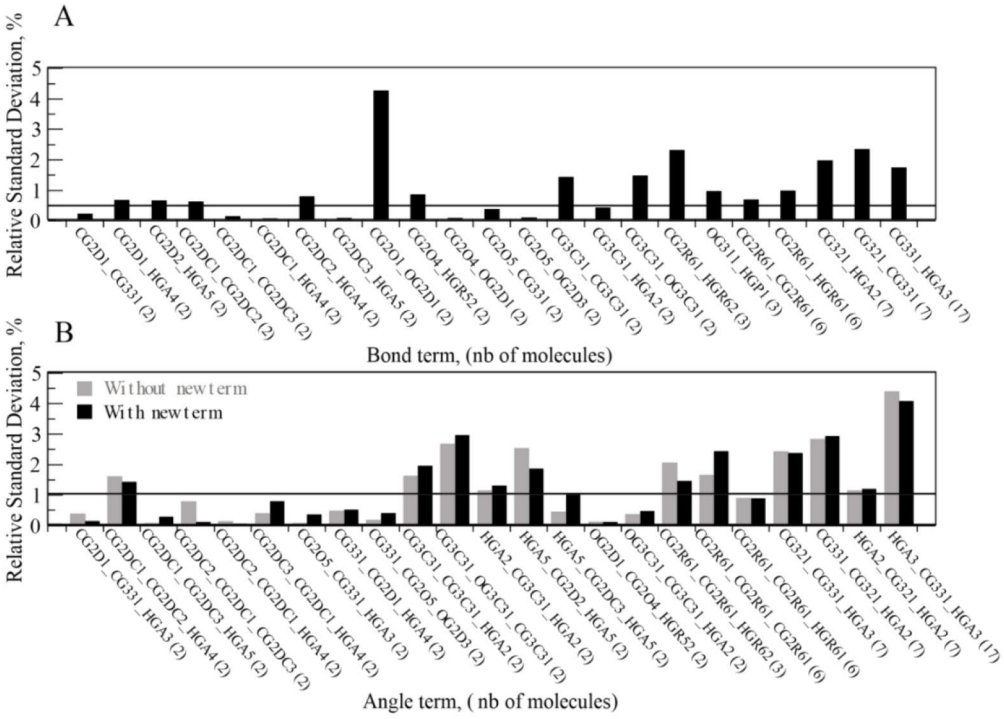


Figure 5.6. Transferability of optimized FF force constants relative to the initial parameters. The relative standard deviation is shown for bond (A) and angle (B) force constants optimized in different molecules. On panel (B) black and gray bars correspond to the results with and without the new angle term given by Equation 9, respectively. The horizontal lines show the average relative SD. The number of molecules sharing the term is given in parenthesis.

Table 5.4 Standard deviation for bonded term parameters optimized in different molecules. The results were averaged over 23 bond terms and 24 angle terms

	Force constants, K_a and K_b		Equilibrium parameters, b_0(Å) and θ_0°	
Optimized terms	Relative SD%	SD	Relative SD%	SD
Bonds	1.0 (1.0)	4.31 (6.7)	0.1 (0.1)	0.0016 (0.0013)
Angles without the new term[a]	2.1 (4.3)	0.63 (0.88)	0.7 (0.6)	0.75 (0.73)
Angles with the new term[b]	2.1 (4.3)	0.65 (0.89)	0.6 (0.6)	0.72 (0.72)

[a,b]For angles, force field parametrization was done without and with the additional deformation-based term included to the cost function; the standard deviations for the computed values are shown in parenthesis; the force constants are in kcal·mol^{-1}·Å^{-2} and kcal·mol^{-1}·rad^{-2} for bond and angle terms, respectively.

The SD averaged over all bond terms for the force constant, K_b is just 4.3 kcal·mol^{-1}·Å^{-2}, with the maximum relative value of 4.5% for the bond term defined for atom types CG2O1-OG2D1. All bond terms except this bond have the relative SD less than 3%. The SD for the equilibrium bond length averaged over all bond terms is just 0.0016 Å. For angle terms, the SD for force constants K_a optimized in different molecules is again very small 0.6 kcal·mol^{-1}·rad^{-2} (the relative SD is 2%). After optimization in different molecules, the equilibrium angles θ_0 are practically identical with SD of 0.8°. Since the force constants and equilibrium parameters deviate insignificantly, the force field parameters optimized by the new method are largely transferable, i.e. can be optimized in one molecule and used for the other molecules, at least as demonstrated for the molecules in the data set.

Normal modes

We further tested the quality of the optimized parameters using normal mode analysis. We note that, typically, for comparison, QM and MM normal modes are sorted based on the magnitude of frequencies.[96,152] However, as shown in Table 5.5, around ~50% of normal modes do not match , i.e. pairs of QM and MM normal modes with the absolute dot product (d) between normalized NMs lower than $d < 0.5$, if they are sorted based on the magnitude of frequencies. Figure 5.7 shows dot product between QM and MM normalized normal modes sorted based on the frequency magnitude for benzene, ethylene, dimethylsulfoxide and butanol. As it can be seen, many corresponding MM normal modes are not in the same order as QM normal modes, and also QM normal modes may have several contributions from MM normal modes. Thus, if QM and MM modes are sorted only based on their frequencies, one may compare QM and MM normal modes which can be even orthogonal. To solve this problem, in this work we establish the correspondence between QM and MM normalized normal modes based on the dot product before the comparison as described in the methods section. With matching normal modes based on the dot product, 100% of QM and MM normal modes have $d > 0.5$, and ~90% have $d > 0.75$ as indicated in Table 5.5. However, if normal modes are sorted based on their frequencies, only ~51% of pairs of QM and MM normal modes have $d > 0.5$, and ~45% have $d > 0.75$.

Table 5.5 Analysis of QM and MM normal modes pairs based on the absolute value of dot product between normalized NM vectors. The results are averaged for 32 molecules set.

	Pairs of normal modes matched based on frequency magnitude			Pairs of normal modes matched based on largest absolute dot product value		
	Proportion of dot product > 0.5	Proportion of dot product > 0.75	Proportion of dot product > 0.9	Proportion of dot product > 0.5	Proportion of dot product > 0.75	Proportion of dot product > 0.9
Initial terms	0.5	0.43	0.33	0.99	0.89	0.64
Bond terms optimized	0.52	0.45	0.35	1	0.89	0.63
Angle terms optimized	0.51	0.43	0.33	1	0.88	0.62
Both bond/angle terms optimized	0.52	0.47	0.38	1	0.9	0.67

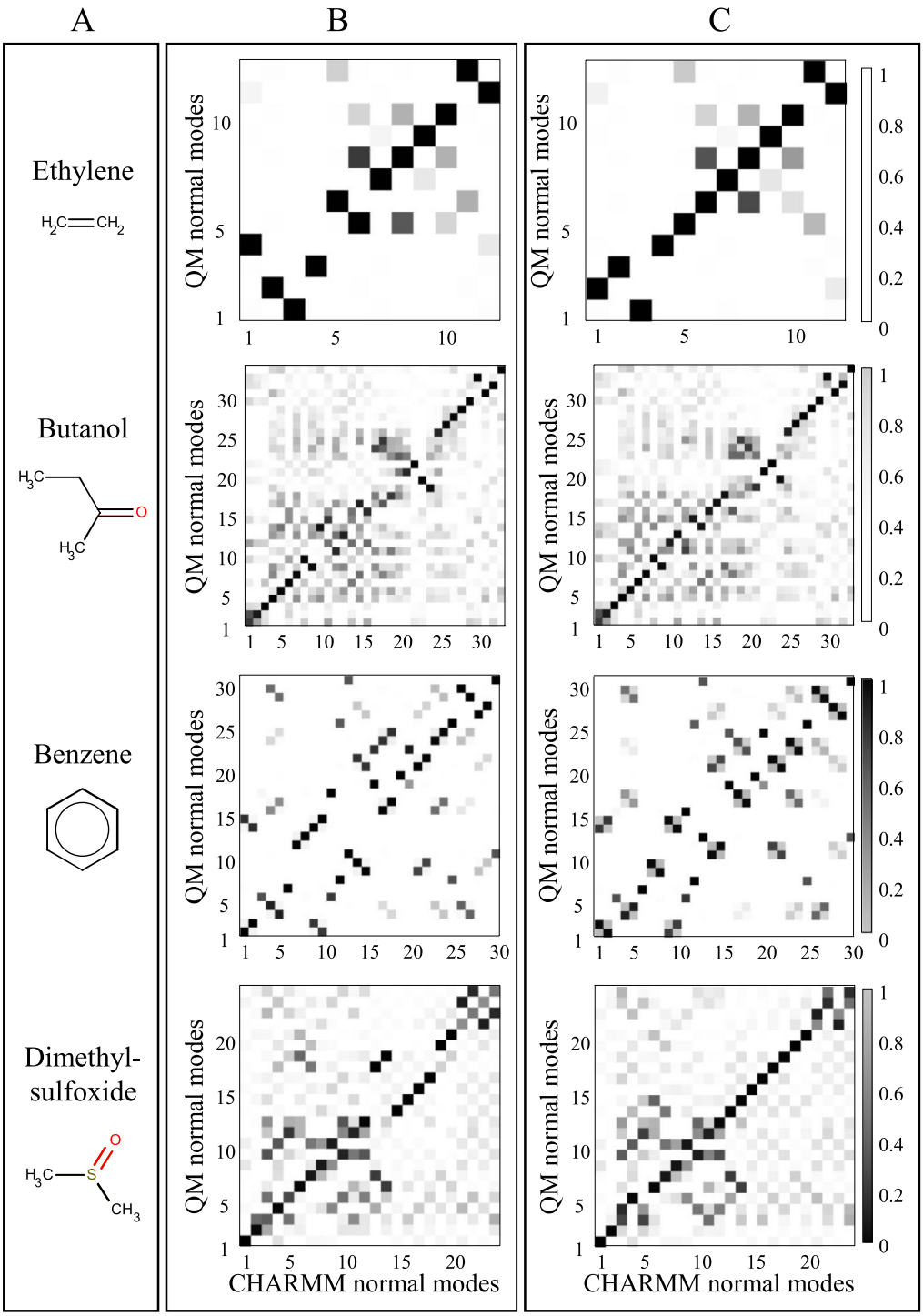


Figure 5.7 Dot product between QM and MM normalized normal modes (NMs) for selected molecules. NMs were computed with the initial CHARMM parameters (B) and with the optimized parameters (C). In (B) and (C), MM and QM normal modes were sorted based on the frequency magnitude. 2D structure representations were prepared with MarvinSketch[129]

To characterize the fit between QM and MM frequencies, Mean Percentage Absolute Error was calculated, after matching QM and MM normal modes based on the dot product. Figure 5.8 demonstrates the improvement of normal modes relative to QM NM for each individual molecule. Table 5.6 gives the mean MPAE for the 32-molecule set. We note that the initial CHARMM parameters were derived to

reproduce normal modes and thus are expected to give very good results relative to the QM normal modes. Indeed, a mean MPAE is 9.5% with the initial CHARMM parameters. With the bond term parameters optimized using the term given by Equation 9, the mean error is lower 8.3%. With both bond and angle terms optimized the MPAE is getting even lower to 6.8%. It should be noted that without matching normal modes based on the dot product, the MPAE is consistently lower with an average of 6.9% with initial CHARMM parameters and with 5.4% for optimized bond and angle terms. However, the optimized parameters still give better results than with the initial parameters.

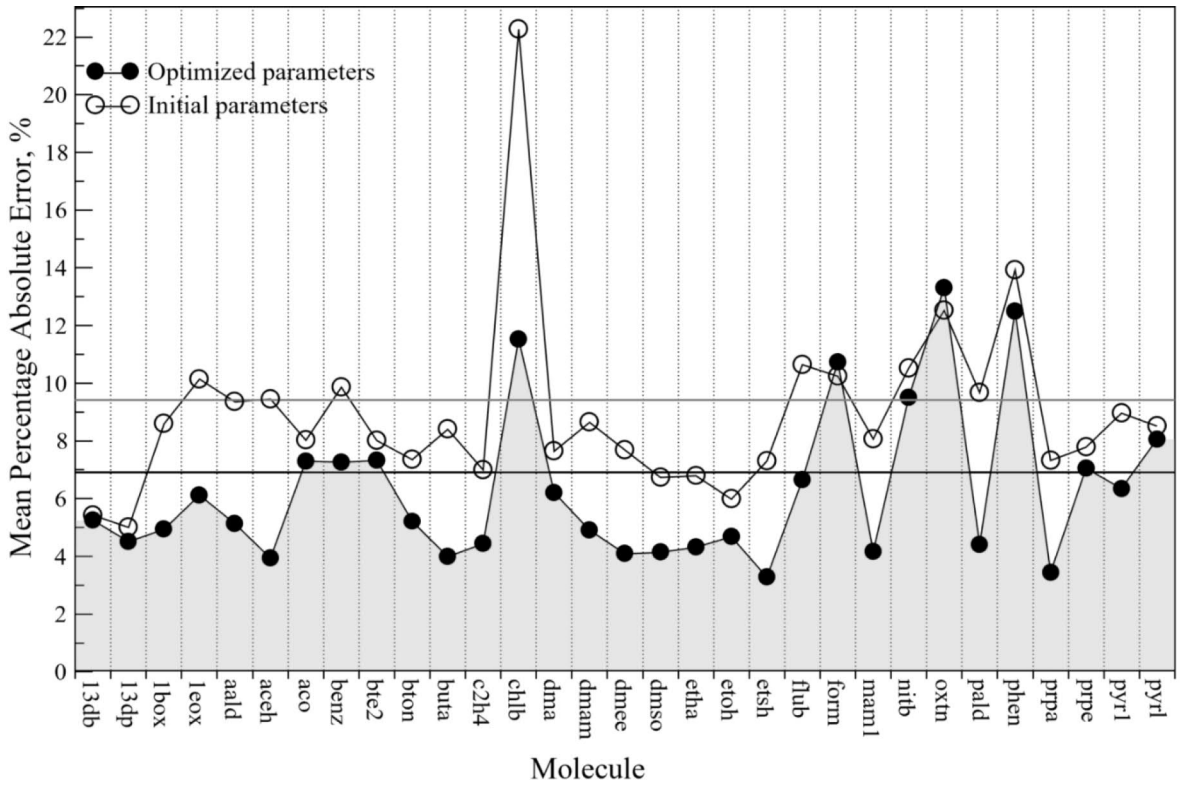


Figure 5.8 Mean Percentage Absolute Error between QM and MM normal modes computed with initial and optimized bonded term parameters. QM and MM normal modes were matched based on the dot product. Average values are represented as horizontal lines. The data are shown for all molecules in the data set, except acetamide (acem) due to its large value, 26.19% and 25.73% with the initial and optimized parameters, respectively.

Table 5.6 Mean Percentage Absolute Error between QM and MM normal mode frequencies averaged over 32 molecules in the test set. The pairs of QM and MM NMs were matched based on the dot product between normalized QM and MM NMs.

	[a]initial	[b]bond terms	[c]angle terms	[d]both bond/angle terms
MPAE[a] %	9.46	8.30	8.04	6.84

[a]MM NMs were obtained with the initial parameters, [b]with the bond term parameters optimized, [c]with the angle term parameters optimized, and with both the bond and angle parameters optimized;

Figure 5.9 compares initial and optimized force constants. For bond terms, as expected, the optimized parameters are in good agreement with the standard CHARMM force constants with the linear correlation coefficient of 90% and RMS deviation of 68.4 kcal·mol^{-1}·Å^{-2} (18%). For angles, the correlation is smaller 69% and the RMS deviation is 14.7 kcal·mol^{-1}·rad^{-2} (31%).

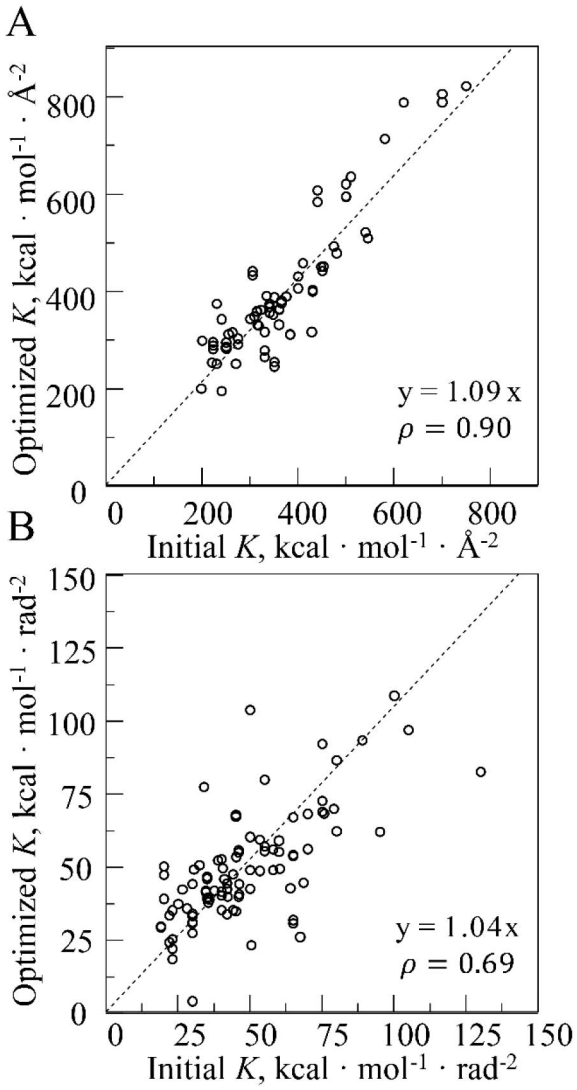


Figure 5.9 Initial versus optimized force constants for (A) bond and (B) angle terms. Force constants are shown for a total of 72 bond terms and 115 angle terms. The linear fit is shown by a dashed line.

CONCLUSION

In the present study we demonstrate that optimization of force field parameters based on comparing QM and MM energies and/or forces of structures leads to suboptimal force constants for bond and angle terms, if structural deviations between QM and MM equilibrium structures are present. The presented results show that conventional parametrization methods based on fitting QM energies and forces are largely balanced toward accuracies in force field equilibrium bond length and angle, while the accuracy in force constants is sacrificed. With the structural deviation present, the optimized force constants cannot adequately describe the QM flexibility of the molecule, as exemplified on several test molecules. Structural deviations always lead to force constants smaller in comparison to those in the absence of such structural deviations, and to a softer MM model relative to the QM model.

To solve this problem, we developed and implemented a new method to derive force field parameters for bond and angle terms. The new method derives force field parameters based on PES scans where a structural deviation between QM and MM structures is allowed. We tested the method on a set of 32 molecules, and the results show that the optimized force field parameters are robust relative to random initial force constants. Starting from five sets of random force constants, we obtained relative SDs of just 0.3% and 1.2% for the bond and angle force constants, respectively. FF parameters derived by the new method are largely transferable, as demonstrated by the low relative SD (< 2%) for equilibrium bond and angle values and force constants for the same terms optimized in different molecules. We further tested the method to reproduce QM normal modes. The results indicate that there is only 45% correspondence between MM and QM normal modes, if they are sorted based on the frequency magnitude for the comparison, underlying the importance of establishing the correct correspondence based on the dot product. Furthermore, without correctly matching QM and MM normal modes, the agreement for normal modes defined by MPAE may appear better (6.9%) than after matching normal modes based on the dot product (9.5%).

Overall, the new method will allow to parametrize molecules with structural deviations present between QM and MM equilibrium structures, common for force fields for small molecules, producing robust and transferable parameters. In future, the method will be extended to derive parameters for dihedral angle terms.

Chapter 6
FORCE FIELD DEVELOPMENT FOR THE ZINC CUSTOM LIBRARY

In this project, which is ongoing at the moment of writing of this manuscript, we are performing the large-scale parametrization of a ligand library. This chapter will start with a brief discussion of the CADD tools based on force field methods; followed by the presentation of available compound libraries; and types of compounds they contain of interest for drug design. Although many compound libraries exist today (Molport[153], ZINC20[154], Enamine[155], AllChem[156], SCUBIDOO[157], SAVI[158], *etc.*) in this work, the choice was made for the ZINC20 library, which will be discussed in detail.

Force field methods can be used at different levels in computationally assisted drug design. Generally, CADD methods for ligand identification can be classified into structure-based (SB) and ligand-based (LB) drug discovery. In this work, we focus on structure-based methods, and in particular physics-based, which will be shortly introduced in what follows. SB methods require information on the 3D structure of the target, which is usually obtained experimentally, but can be also obtained through modelling, if structures of homologous proteins are present. In recent years, powerful methods have appeared for structure modelling based on the artificial intelligence, with arguably the most known AlphaFold,[159] making it possible to apply SB methods to a much wider range of protein targets with unknown structures. Once the 3D structure of the target is available, it is possible to apply docking/scoring methods for virtual screening. Docking consists in positioning the ligand in the binding site of the target and finding the optimal pose for highest binding affinity, while scoring consists in evaluating binding properties of ligands with binding poses obtained from docking. Typically, a set of compounds or a library is tried against the target in the virtual screening. Scoring consists in evaluating the affinity of ligands to the target by calculating *(scoring) functions*, which use three main approaches: knowledge-based, empirical, and force field-based functions.[160] The notion of scoring can be also used for selecting different binding poses and different scoring functions can be used at different steps (docking, post re-scoring etc.; see below).

Force field-based scoring functions estimate the binding energy using force fields by summing the contributions of bonded and non-bonded terms, and can be supplemented with implicit solvent models (MM/PBSA or MM/GBSA).[161] Table 6.1 provides a list of several scoring functions implemented in the most frequently used molecular docking programs.[162] Docking methods can be reasonably successful in predicting the conformation of the ligand within the target binding site. However, the difficulty is to reproduce the absolute binding free energy, and relative binding affinity between significantly different ligands.[162] The problem is due to the fact that ligand binding represent a large cancelation of different contributions due to solvation/desolvation, favourable interactions with the receptor, and loss/gain of entropy. All these contributions should be estimated with a significant accuracy. To this end, FF-based scoring methods, which in principle take all details of the atomistic structure, can bring the accuracy in solving this problem. Current force fields for small molecules, such as the CHARMM CGenFF reached the level, where parameters are available for thousands of small compounds. However, to be applied to real screening studies, typically done with libraries containing sub-million and million molecules, FF models should be further developed.

Table 6.1. Frequently used scoring functions adapted from Ferreira *et al* (2015)[162]

Force-Field-Based	Empirical	Knowledge-Based
Dock	AutoDock	SMoG
AutoDock	GlideScore	DrugScore
GoldScore	ChemScore	PMF_Score
ICM	X_Score	MotifScore
LigandFit	F_Score	RF_Score
Molegro Virtual Docker	Fresno	PESD_SVM
SYBYL_G-score	SCORE	PoseScore
SYBYL_D-score	LUDI	
MedusaScore	SFCscore	
	HYDE	
	LigScore	

It should be noted in the context of FFs in CADD, recent methods based on MD simulations appeared as an alternative to rigid-body docking. In these methods, MD simulations are typically performed with the target immersed into water and small molecules. Typically, to simulate more frequently association/dissociation processes of such small molecules to/from the target the concentration of small molecules has to be large. These small molecules can be from the library, or just

fragments, as it is done, for example, in Site-Identification by Ligand Competitive Saturation (SILCS)[163]. SILCS relies on classical molecular dynamics simulations to model competitive binding of small molecules to obtain the affinity pattern of target macromolecules (proteins) for chemically diverse functional groups

Compound databases are widely used in CADD.[164] Ligand databases may contain different properties, such as 1D, 2D and/or 3D structures of the compounds; chemical and/or physical properties; information about their synthesis reactions; information about their reactions in biological pathways; potential biological targets and importantly vendors for a compound. Most of the times, the databases also provide online tools to perform searches amongst their compounds. Table 6.2 lists some of the largest databases. For this work, we chose the ZINC library that can be freely accessed with all ligands available for purchase.[154]

Sampling the drug-likeness of ligands is a complicated task given the immense number of possible unique organic molecules, referred to as the chemical space. The covered chemical space depends on the size of ligands. For instance, molecules containing 17 heavy atoms (only C, N, O, S and halogens) can form up to 166.4 billion possible organic molecules.[165] However, the space of possible compounds that can be actually synthesized by organic chemists is very limited. For example, chemical abstract service (CAS) registry that documents every chemical substance described in the open scientific literature contains 196 million organic and inorganic substances in total.[166]

Table 6.2 Available compound libraries adapted from van Hilten *et. al* (2019)[164]

Name	Size	Compound availability	Comments
GDB-17[167]	$\sim 166 \times 10^9$	-	freely accessible
ZINC20[154]	$\sim 500 \times 10^6$	purchasable	freely accessible
Enamine REAL[155]	$> 300 \times 10^6$	purchasable	commercial
SCUIBIDOO[157]	$\sim 21 \times 10^6$	synthesizable	freely accessible
CHIPMUNK[168]	$\sim 95 \times 10^6$	synthesizable	freely accessible
AllChem[169]	$> 10^{20}$	synthesizable	commercial
PLC[170]	$\sim 10^{11}$	synthesizable	commercial

ZINC20

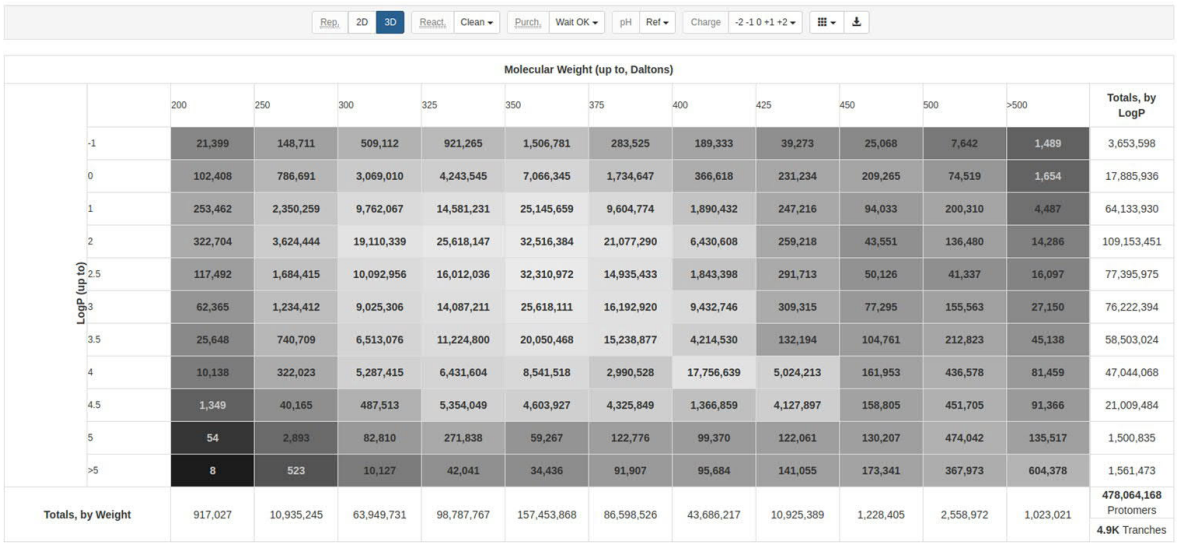
ZINC (an acronym for "ZINC is not commercial") is a publicly accessible chemical database that contains commercially available and annotated compounds, and is very popular for ligand discovery purposes.[154] It is updated and curated on a regular basis. ZINC was developed by John Irwin in the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco,[171] with the most recent version, ZINC20.[154] ZINC library represents a broad spectrum of chemical space and is widely used by the scientific community. The ZINC website is extensively used by thousands of investigators (monthly), and many terabytes of data are downloaded each week.[154]

ZINC was conceived to provide subsets of molecules with variable properties such as functional groups, molecular weight, and calculated polarity (logP) that should be easy to create and manipulate. The database supports multiple protonation models, tautomeric forms, stereochemistries, suppliers, and 3D conformational sampling. Other databases and libraries, such as HMDB[158], ChEMBL[172], and DrugBank[173] were used to obtain the biological annotations, the identification of molecules as metabolites, drugs, or natural products and the identification of molecules as ligands for particular proteins and processes.[174] The above-mentioned properties of ligands can be used to perform queries in the library. In addition, structural characteristics of the molecules can be used to perform: whole-molecule search in which molecules that most resemble the entire query are prioritized;

substructure search in which the molecules that contain the entire query molecule are identified; and pattern search in which molecules containing specified molecular pattern(s) are selected.[154] For more general queries, ZINC20 provides the Tranche Browser shown in figure 6.1. The Tranche Browser divides the physical property space of the ligands into 121 tranches based on two properties in 11 bins each: the horizontal axis is molecular weight, and the vertical axis is logP. The Tranche Browser allows the selection of the characteristics of the database subset required and then to download it, in SMILES or 3D formats.[174] Additional options like purchasability, reactivity, the different protonation states in function of pH and the net charge of the ligand are proposed. However, the last are possible only for molecules with a 3D available structure which are also called protomers.[174]

Properties for the protomers in ZINC20 are obtained with ChemAxon package that determine protonation states and tautomers at physiologically relevant pH in three pH tranches. These pH tranches covered reference pH (6.4 to 8.4), high pH (8.4 to 9.0), and low pH (5.8 to 6.4). Each protomer was rendered into 3D using the ChemAxon package (ChemAxon, Budapest, Hungary) and conformationally sampled using Omega100 (OpenEye Scientific Software, Santa Fe NM). Files formats for docking are accessible under the multiple formats (mol2, sdf, and pdbqt).[154]

Figure 6.1. Extract from Zinc database accessed on 07/11/2021 showing tranches of molecules https://zinc20.docking.org/



CUSTOM LIBRARY

We created a selection of molecules for further parametrization using the following criteria: potential molecules with drug-like properties; a manageable size of the library for performing extensive *ab initio* calculations; and molecules representative of wide regions of the chemical space.

DRUG-LIKE PROPERTIES

The goal was to create a library of molecules with "general" drug-like properties. As there is no such universal gold standard measure to contour the drug-like chemical space, we could apply the popular selection methods based on the "Rule of 5" of Lipinski, the work of Veber based around PSA, the "GSK 4/400" for lead-like molecules or the "Rule of 3" for fragment molecules.[175] We decided to use the available information for compounds that are actually validated drugs or that are still in clinical trials.

The fundamental building blocks of approved drugs are ring systems. Moreover, rings are of great interest for modern medicinal chemists, since they play an important role in molecular properties such as the electronic distribution, three dimensionality, and scaffold rigidity.[175] There are approximately 450,000 unique ring systems derived from 2.24 billion molecules currently available in synthesized chemical space, and molecules in clinical trials utilize only 0.1% of this available pool.[176] The recent work of Jonathan Shearer and al. shows that current drug approved space comprises only 378 ring systems and each year and, on average, only 33% of new drugs contain one new ring system. Furthermore, approximately 50% of the novel ring systems entering clinical trials are systematic changes of up to two atoms on existing drug resulting in a set of 3,902 future clinical trial ring systems.[176]

SELECTING LIGANDS

In order to obtain a set of molecules for performing extensive *ab initio* calculations we applied the following criteria:

i. 3D geometry available;
ii. available for sale directly or on demand;
iii. a predominant form at pH 7;
iv. not reactive;
v. molecular weights between 150 – 250 Da;
vi. contain rings;
vii. no more than one rotatable bond between heavy atoms; however, there can be any number of rotatable bonds with hydrogens;
viii. contain only the most abondant elements: C, N, O, H, S, F;

Although ZINC20 is supposed to perform fast research in its database by applying specific criteria, like the ones listed above, it failed to provide such a custom library (the source of the error is unknown for us). With the help of ZINC20 tranches browser only criteria *i* to *iv* could be used to obtain an initial library of 75.8 billion molecules that have a molecular weight lower than 300 Da. Next, criteria *v, vii* and *viii* were successfully applied with the help of OpenBabel[101] software. Molecules between 150 and 250 Da with one or no rotatable dihedrals contain by default ring substructures, so criterion *vi* was also applied. The custom library applying the described selection contains 285,041 molecules.

The availability of 3D geometries is necessary to obtain correct structures of the molecules with a unique form that is a specific stereoisomer, tautomer and has a single protonation form. This information would be missing if 2D formats were used. Although ZINC is supposed to have commercially available compounds, some of them cannot be provided immediately upon request. However, since we wanted to include an extended set of molecules, compounds available directly for sale or on demand were considered. Following the experience with nonstandard amino acids optimization, we preferred directly obtaining the predominant protonation at physiological pH between 6.4 and 8.4. Also, to limit the amount of QM computations we decided to have relatively small molecules containing rings and thus the molecular weight limitation. Also, since rotatable dihedrals require 6 times more QM target data compared to any other bonded term, we wanted to limit their number based on limited number of rotatable bonds. We included only molecules containing

the atoms C, N, O, H, S, F as these are the atoms generally present in drugs. P and other halogens except F were excluded because typically they are longer to optimize. Finally, only non-reactive molecules were included, as the CGenFF force field is not adapted for radicals.

To parametrize selected 285,041 molecules, representative compounds were selected and created as follows. First, we generated initial parameters with CGenFF program[105]. Secondly, we produced a list of ZINC molecules to represent the entire chemical space in the 285,041 molecules based on bonded terms sorting. Finally, when possible, we fragmented the molecules in this list, and we repeated the second step to obtain a reduced size library that contains ZINC molecules as well as fragments. The reduced library contained all the parameters present in the 285,041 molecules.

The mol2 structures for the 285,041 molecules were downloaded from ZINC20 database and atom names were conserved with hydrogens already present. CGenFF initial parametrization was done in the MacKerell lab at the University of Maryland, where a python pipeline using CGenFF program was set up locally. However, not all molecules were processed successfully, that resulted in 0.36% data loss as shown in Table 6.3.

Table 6.3 Pipeline statistics of concurrent CGenFF processing

Pipeline metric	Value
CGenFF Compounds Passed	274 736
CGenFF Compounds Failed	10 305
Total Compounds Processed	285 041
Data Loss	0.36%

BONDED TERMS-BASED SORTING

The initial parameters and topologies were generated with the CGenFF program[105] from the 3D geometry and the covalent structure for each molecules. Additional to initial parameter CGenFF program also provided a penalty term. Only parameters with penalty >10 were considered for optimization. A lower penalty indicates that the substructure is already present in the CGenFF force field.

The CGenFF processed compounds contained in total 180,992 bonded terms with high penalties that need optimization and 123,992 bonded terms with low penalty that can be adapted as they are. When we eliminated the duplicates, we obtained 34,500 (73.6%) unknown terms and 12,363 (26.4%) known terms. Known terms have low penalties and unknown terms have high penalties. The proportion of molecules that had parameters that require optimization according to our criteria is 11.8% (32,500 molecules).

The list of unique parameters was created from all parameters by removing repeated bonded terms and the information on molecules containing particular bonded terms was stored. We obtained two such lists, one for parameters with high penalties and one for low penalty parameters. Next, we used the list with unique unknown parameters to determine the lowest number of molecules containing all of them and also the order of optimization for these molecules so each bonded term would be optimized only once. Moreover, the order of optimization would be organized by groups. First group to be optimized would be constituted by molecules that do not share any bonded terms between them. Second group will also be made by molecules that don't share any bonded terms between them inside the group. However, the second group can share parameters with the first group. So, these shared terms are optimized in group I and then adapted in group II. In group II, a new set of terms is optimized. The number of groups was adjusted to include all bonded terms.

Molecules in a group were selected in three steps. First, neutral molecules were prioritized. Second, among the molecules of the same charge, compounds with the largest number of unknown bonded parameters had higher priority. Finally, among molecules with the same charge and the same number of unknown parameters, compounds having a smaller number of atoms were selected. To prepare the second group, the terms included in the first list were removed from the initial list of unique bonded parameters. The iteration of the selection described above to create the first group of molecules is repeated giving a second group of molecules. The

iteration is repeated until the list of unique bonded parameters does not contain any unknown parameter. Figure 6.2 summarizes the entire process.

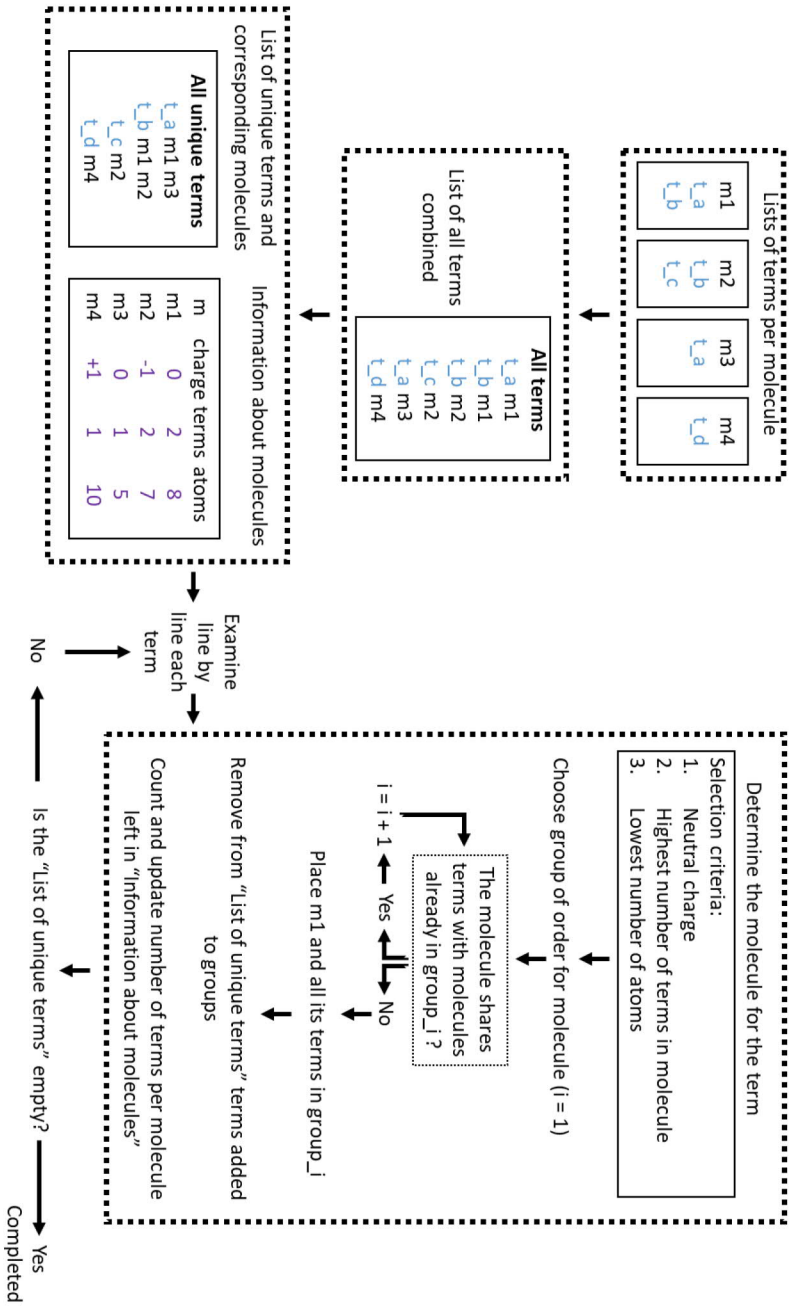


Figure 6.2 Schematic representation of the iterative creation of groups of order containing molecules to be submitted for force field optimization. Molecules are named with lower letter "m", while the terms are named "t_ ". All terms are combined and sorted to obtain a list of unique terms. Using the list of unique terms and information about molecules, iterations for group selection can be started. In each iteration, a term from the unique list is examined, here t_a being the first term. If multiple molecules share this term, the neutral molecule with a higher number of terms and lower number of atoms is selected for inclusion in the first group. t_a is present in m1 and m3. According to selection criterion, m1 is retained. In the first iteration, all terms of the selected molecule m1, are added to group 1 and molecules sharing terms with m1, like m2, can't be in the same group. Then, all terms contained by m1 are removed from unique list of terms and information about molecules is updated. In the second iteration, the next term is examined: t_c. Since t_c is only present in m2, m2 can directly be added to group 2. The lists are updated, and the iteration is continued until the list of unique terms becomes empty. In the example, three iterations were sufficient.

We applied different orders to select representative molecules in groups to optimize bonded terms, based on number of atoms, the net charge of the molecule and number of bonded terms that need parameters to optimize. The results of the obtained 6 different lists are summarized in Table 6.4. All six lists have similar lengths comprised between 6,188 and 6,756 molecules, similar average number of atoms per molecule varying from 26.5 to 28.4, and comparable number of charged molecules going from 3,990 to 4,412 molecules per list. During the parameterization process it is preferable to have neutral molecules, which let us choose the list with 6,497 molecules with priority inclusion criteria being charge more important than number of terms, number of terms being more important than number of atoms (I. Charge II. Terms III. Atoms). The selected list has 6,497 molecules with only 3,990 being charged. On average there are 25 unknown terms and 28.1 atoms per molecule in this list. More importantly, 34.9 % of the list contains molecules that share terms with more than 100 other molecules and 86.7 % of the list share terms with at least 10 other molecules from the 200,000 molecules list.

Table 6.4. Different selection methods by varying the priority of the inclusion in the list. The 3 criteria for inclusion were the charge, the number of atoms and the number of unknown terms contained by the molecule which gave 6 different lists. The priority of inclusion is indicated by roman numbers.

Type of list	Number of molecules	Average number of atoms per molecule	Minimum number of atoms per molecule	Maximum number of atoms per molecule	Number of charged molecules (%)	Average number of unknown terms per molecule	Proportion of molecules sharing terms with more than 10 molecules (%)	Proportion of molecules sharing terms with more than 50 molecules (%)	Proportion of molecules sharing terms with more than 100 molecules (%)
All molecules with unknown terms	32410	29.7	10	49	2,3832 (73.53%)	18.3			
I. Charge II. Atoms III. Terms	6,762	27.3	10	48	4,147 (61.33%)	24.6	86.7	57.6	34.9
I. Charge II. Terms III. Atoms	6,497	28.1	11	48	3,990 (61.58%)	25	86.6	57.5	35.1
I. Atoms II. Charge III. Terms	6,756	26.5	10	48	4,401 (65.14%)	24.4	86.7	56.4	33.9
I. Atoms II. Terms III. Charge	6,721	26.5	10	48	4,412 (65.65%)	24.4	86.7	56.4	33.9
I. Terms II. Atoms III. Charge	6,174	28.3	11	48	4,099 (66.39%)	25.2	86.5	56.8	34.1
I. Terms II. Charge III. Atoms	6,188	28.4	11	48	4,080 (65.93%)	25.2	86.5	56.8	34

When possible, we added additional molecules containing neutral functional groups that initially were ionized in the original ZINC compound. Such compounds were renamed with the ZINC ID plus "n" letter. The neutralized functional groups by protonation/deprotonation included hydroxyls, thiols, amines, and imines. In total, 890 neutral compounds were added resulting in a list of 7387 compounds to optimize.

Our library is made up of ring containing molecules. Fragmenting a ring is not optimal for bonded terms transferability. However, rings separated by a rotatable bond can easily be used as fragments. We developed a Python3 script to automatically detect such bonds, cut these bonds, and replace the eliminated ring by a methyl group as shown in Figure 6.3. However, rings connected through the rotatable bond to a nitrogen atom were not accepted as fragments because of the planar conformation the nitrogen has in a ring that cannot be approximated by a simple functional group.

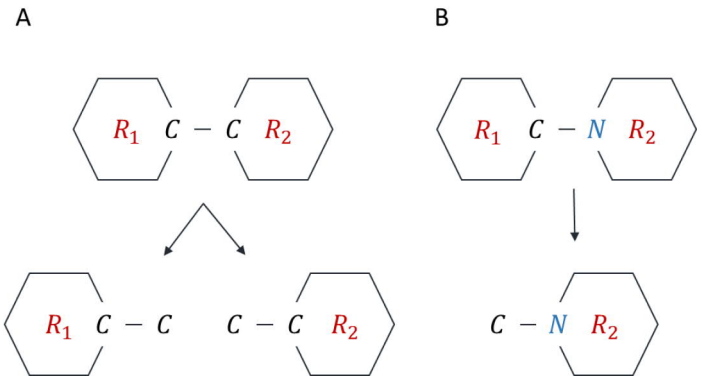


Figure 6.3 Fragmentation of model compounds containing two or more ring structures. (A) When rings R1 and R2 are connected by a rotatable bond containing carbon terminal atoms, the cleavage results in the creation of two rings containing a methyl at the cleavage point. (B) When the rotatable bond contains a carbon and a nitrogen atom, only the ring on the nitrogen side (R2) will be used as a small compound with a methyl at the separation of point.

The Python script used RDKit module[177] and PyMOL[107] program python compatible interface. RDKit is an open-source toolkit for cheminformatics. It has multiple functionalities, but we were interested in 2D and 3D molecular operations.

In total, there were 2,997 soft bonds between rings that after cleavage resulted in 4,313 fragments. These rings are not representative for the totality of chemical set in the library, so optimization of complete molecules was also performed. In total, there are 6,688 soft rings that can have different puckering states and 4,110 rigid rings for fragments and complete molecules combined.

Next, we established the order of compound optimization only for rigid rings. First, we applied the sorting method based on bonded terms on the fragment rigid rings that were reduced to 387 to be optimized in seven groups. The bonded terms present in the 387 fragments were excluded from the list of bonded terms of complete molecules made only of rigid rings. The same method was applied to complete molecules which resulted in 8 groups for optimization with a total of 1,955 molecules. So, the final list for rigid rings was reduced from 4,110 to a total of 2,342 molecules.

In the process of minimizing the library size, it was important to maintain the chemical diversity that can actually be useful for the scientific community. Our collaborators in the MacKerell lab at the University of Maryland created a tool called GlobalChem[178] that identifies whether the functional groups in a library are present in important collections of compounds. Our library contains an important number of functional groups that are already used by the scientific community as shown in Table 6.5. For instance, in our library there are 8,311 functions found in the common ring scaffolds in FDA approved drugs[175]. Our library is not only representative of existing chemical space in the literature, but also is very rich in substructures absent from CGenFF force field as indicated by the abondance of high penalty terms present.

Table 6.5 Availability of the selected chemical space in the literature

Chemical List	Functional Group Match Count
Amino Acids	20,923
Organic Solvents[179]	20,722
Open Smiles[180]	25,069
IUPAC Blue Book Common Rings[181]	5856
Common Heterocyclic Rings in Phase 2[182]	10,585
Rings in Drugs[183]	8,311
Privileged Scaffolds[184]	1798
Common Warheads[185]	776
Common Polymer Repeating Units[181]	1,235

Through the devised protocol explained above, we selected 285,041 molecules from ZINC20 database for their potential drug-like properties, representing a wide part of chemical space and absent from CGenFF. We also applied the necessary measures to be able to perform the force field optimization for this custom library in a limited amount of time and resources. Using our method based on bonded terms sorting we reduced the size of the library to less than 8,000 molecules. Additionally, we created fragments, also called model compounds, to decrease the workload for QM computations.

For the force field development, we automatized the identification of soft dihedrals and fragmentation. The process of optimization is currently performing using high performance computers available to our lab.

Chapter 7
CONCLUSIONS AND PERSPECTIVES

The main goal of my PhD work was to significantly extend the coverage of the chemical space by the CHARMM FF and develop new tools for the improved FF development. To this end, during the first 18 months of my thesis, I extended the CHAMM FF and CGenFF to a large set of nonstandard amino acids frequently present in PDB structures and to a set of small molecules, respectively. The selected nonstandard amino acids are of both natural and artificial origin and present chemical modifications at the level of sidechain and/or at the level of the backbone group. For the force field development, I created a set of small compounds representing different functional groups in nonstandard amino acids, including different dipeptides and tripeptides to account for the peptide bond. Partial charges, bond, valence angle, dihedral and improper torsion terms were considered for optimization, which was performed according to the standard CHARMM method to balance interactions with nonstandard amino acids and with other components of the simulation system, described by the FF developed in this work and by the standard CHARMM FF, respectively. The optimization of intra- and inter-molecular terms largely relied on fitting the target *ab initio* data, mainly due to the absence of available experimental data, and demonstrate agreement with QM and experimental data similar to the one achieved for the standard C36 FF and CGenFF. In particular, the protein models in the MD simulations mostly fluctuated around the experimental structures as demonstrated by different criteria which include amongst others important dihedral torsions for backbone modified, and sidechain-associated χ_1 and χ_2 for sidechain modified amino acids. For both types of chemical variants of amino acids, sidechain or backbone modified, it was difficult to integrate the force field into the general framework of the standard CHARMM FF requiring adjustments and compromises for the general protocol. The main drawback of the current implementation is that the nonstandard amino acids are represented as a mixture of atom types and corresponding parameters from both C36m and CGenFF. It should be noted that C36m and CGenFF largely overlap, as many parameters used in C36m were adapted for CGenFF. Obviously, in the current state of the development it represents a compromise, as many parameters are simply repeated in C36m and

CGenFF, and in future these FFs will be united, using uniform atom types for standard, nonstandard amino acids, and small molecules.

The CMAP term, considered only for the sidechain-modified amino acids in my work, can be specifically developed for backbone-modified amino acids if a better treatment for peptide backbone conformation is needed in future. Force field development can be continued to other nonstandard amino acids, which were considered as "rare" in this work, and for this reason, not included to the FF development. In addition, other types of modifications can be considered, which cannot be classified as *amino acids,* such as imines. It should be also noted that in the current CHARMM force field, selenium-containing groups are treated as sulfur-analogs (for example seleno-cysteine and seleno-methionine are modeled as cysteine and methionine, respectively), with the same parameters used for both atoms (*S* and *Se*). The main difficulty (our unpublished results and personal communication with Alex MacKerell) is that for the sulfur-containing molecules, including cysteine and methionine, the standard additive form of the CHARMM FF does not reproduce interactions with solvent well at all interaction distances in comparison to QM data, suggesting that a different FF functional form is required. Finally, the validation, or better to define as an illustration, of the force field developed for the nonstandard amino acids was performed by MD simulations of twenty selected different protein systems. Similar to the standard CHARMM FF, where the validation and corrections are mostly done *a posteriori* through a large body of subsequent studies, as exemplified by a recent revision of the standard CHARMM FF needed to better simulate properties of intrinsically disordered proteins, the validation of the FF developed in my work will be done in future studies.

Based on the FF development for the nonstandard amino acids, we made a trivial observation that the empirical model not always reproduces the QM geometry of molecules in vacuum, which can be due to the limited FF functional form, such as the absence of higher order contributions to bonded terms; or due to the adaptation of existing parameters, i.e. not specifically optimized for this molecule, which is done to avoid the need to optimize all parameters for novel molecules. As we demonstrated, while not being important for applications in a general case, this structural deviation between QM and FF structures leads to suboptimal force constants in the FF development. To address this issue, we developed a new method for the FF development of bond and valence angle terms, which does not require that QM and FF geometries coincide in the conformational space. Similar to the parametrization

of soft dihedral angles, this method relies on reproducing the target QM PE surfaces, produced by one-dimensional adiabatic scans in PES. The novelty of the method is that it allows a structural deviation between QM and MM structures as long as agreement for PE surfaces is obtained. The method was extensively tested on a set of 32 small molecules with available CHARMM parameters. We showed that the new method produces stable parameters regardless the initial parameters. As our tests show, the parameters are also transferable, i.e. which can be optimized in one molecule and used in other similar chemical contexts. We also demonstrated that the FF model produced by the new method is described by a better or equal agreement with QM normal modes, than the standard CHARMM FF, demonstrating the quality of the optimized force constants. As a perspective, the method can be further developed for rotatable dihedral angles; however, changes to the methods are required as PES associated with soft dihedrals may have several minima. The optimization tool will be released as a standalone program for the scientific community.

In the final part of my thesis, I focused on extending the CGenFF force field for the ZINC20 library of drug-like molecules. I devised a general and largely automated protocol to select representative molecules from the library based on principles that the selected molecules should significantly cover the chemical space of drug-like molecules and also increase the coverage of CGenFF for new molecules. Representative molecules were further selected for FF optimization based on the principle that a molecule should contain a maximum number of missing parameters, in addition to other criteria, mostly for the ease and convenience of QM calculations. By applying this protocol from the initial 285,041 molecules comprising ring substructures and a few to no rotatable bonds, the library was reduced to 7,387 to represent the entire chemical space of the initial library and to be further optimized. The extraction of properties like rotatable dihedrals, rotatable bonds separating rings, and types of rings was perform in an automated manner. Currently, the optimization of parameters is in progress. QM target data were generated for water interactions, dipole moment and electrostatic potential required for partial charge optimization. QM PES scans needed for bonded terms optimization were generated for molecules containing rings that are planar (rigid). As a perspective, QM multidimensional PES scans will be performed for non-planar rings (soft), which can exist in different puckering states. The validation part of this project will be done by

MD simulation of small molecule crystal structures available from the Cambridge Crystallographic Data Centre (CCDC).[186]

The number of available chemicals is increasing exponentially as witnessed by the geometrical growth of small molecule libraries of commercially available compounds (for example ZINC20 currently contains several hundreds of 10^6 molecules). This growth is accompanied by the recent development of force fields for small molecules, mostly during the last decade: CGenFF[95,99,100], OpenFF initiative[96] etc. Force field based *in silico* tools have gained a popularity especially in the domain of drug design. However, available FFs do not contain parameters for all molecules, including commercially available compounds, underlining the need to expand the force field to accommodate additional molecular variations. The present thesis provides such work for the CHARMM force field and CGenFF extension to a large set of molecules for a wide chemical space. It also presents a new optimization method for improved parameters that maintain compatibility with existing force field while assuring generation of robust and transferable parameters.

REFERENCES

(1) Machado, M. R.; Zeida, A.; Darré, L.; Pantano, S. From Quantum to Subcellular Scales: Multi-Scale Simulation Approaches and the SIRAH Force Field. *Interface Focus* **2019**, *9* (3), 20180085. https://doi.org/10.1098/rsfs.2018.0085.

(2) Korkut, A.; Hendrickson, W. A. A Force Field for Virtual Atom Molecular Mechanics of Proteins. *Proc. Natl. Acad. Sci.* **2009**, *106* (37), 15667–15672. https://doi.org/10.1073/pnas.0907674106.

(3) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750–760. https://doi.org/10.1021/jp036508g.

(4) Bond, P. J.; Sansom, M. S. P. Insertion and Assembly of Membrane Proteins via Simulation. *J. Am. Chem. Soc.* **2006**, *128* (8), 2697–2704. https://doi.org/10.1021/ja0569104.

(5) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666. https://doi.org/10.1021/ja00214a001.

(6) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106* (3), 765–784. https://doi.org/10.1021/ja00315a051.

(7) Reiher, W. E. Theoretical Studies of Hydrogen Bonding, 1985.

(8) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676. https://doi.org/10.1002/jcc.20090.

(9) Jorgensen, W. L. Optimized Intermolecular Potential Functions for Liquid Alcohols. *J. Phys. Chem.* **1986**, *90* (7), 1276–1284. https://doi.org/10.1021/j100398a015.

(10) 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. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5.

(11) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. https://doi.org/10.1021/ja00051a040.

(12) Cieplak, P.; Caldwell, J.; Kollman, P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J. Comput. Chem.* **2001**, *22* (10), 1048–1057. https://doi.org/10.1002/jcc.1065.

(13) Dang, L. X.; Rice, J. E.; Caldwell, J.; Kollman, P. A. Ion Solvation in Polarizable Water: Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1991**, *113* (7), 2481–2486. https://doi.org/10.1021/ja00007a021.

(14) Caldwell, J. W.; Kollman, P. A. Cation-.Pi. Interactions: Nonadditive Effects Are Critical in Their Accurate Representation. *J. Am. Chem. Soc.* **1995**, *117* (14), 4177–4178. https://doi.org/10.1021/ja00119a037.

(15) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force

Field for Simulations of Proteins and Peptides. *J. Comput. Chem.* **2006**, *27* (6), 781–790. https://doi.org/10.1002/jcc.20386.

(16) Grossfield, A.; Ren, P.; Ponder, J. W. Ion Solvation Thermodynamics from Simulation with a Polarizable Force Field. *J. Am. Chem. Soc.* **2003**, *125* (50), 15671–15682. https://doi.org/10.1021/ja037005r.

(17) Wu, J. C.; Piquemal, J.-P.; Chaudret, R.; Reinhardt, P.; Ren, P. Polarizable Molecular Dynamics Simulation of Zn(II) in Water Using the AMOEBA Force Field. *J. Chem. Theory Comput.* **2010**, *6* (7), 2059–2070. https://doi.org/10.1021/ct100091j.

(18) Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7* (10), 3143–3161. https://doi.org/10.1021/ct200304d.

(19) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9* (9), 4046–4063. https://doi.org/10.1021/ct4003702.

(20) Mu, X.; Wang, Q.; Wang, L.-P.; Fried, S. D.; Piquemal, J.-P.; Dalby, K. N.; Ren, P. Modeling Organochlorine Compounds and the σ-Hole Effect Using a Polarizable Multipole Force Field. *J. Phys. Chem. B* **2014**, *118* (24), 6456–6465. https://doi.org/10.1021/jp411671a.

(21) Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **2018**, *14* (4), 2084–2108. https://doi.org/10.1021/acs.jctc.7b01169.

(22) Patel, S.; Brooks, C. L. CHARMM Fluctuating Charge Force Field for Proteins: I Parameterization and Application to Bulk Organic Liquid Simulations. *J. Comput. Chem.* **2004**, *25* (1), 1–16. https://doi.org/10.1002/jcc.10355.

(23) Patel, S.; Mackerell, A. D.; Brooks, C. L. CHARMM Fluctuating Charge Force Field for Proteins: II Protein/Solvent Properties from Molecular Dynamics Simulations Using a Nonadditive Electrostatic Model. *J. Comput. Chem.* **2004**, *25* (12), 1504–1514. https://doi.org/10.1002/jcc.20077.

(24) Zhong, Y.; Bauer, B. A.; Patel, S. Solvation Properties of N-Acetyl-ß-Glucosamine: Molecular Dynamics Study Incorporating Electrostatic Polarization. *J. Comput. Chem.* **2011**, *32* (16), 3339–3353. https://doi.org/10.1002/jcc.21873.

(25) Lucas, T. R.; Bauer, B. A.; Patel, S. Charge Equilibration Force Fields for Molecular Dynamics Simulations of Lipids, Bilayers, and Integral Membrane Protein Systems. *Biochim. Biophys. Acta BBA - Biomembr.* **2012**, *1818* (2), 318–329. https://doi.org/10.1016/j.bbamem.2011.09.016.

(26) Ou, S.; Patel, S. Temperature Dependence and Energetics of Single Ions at the Aqueous Liquid–Vapor Interface. *J. Phys. Chem. B* **2013**, *117* (21), 6512–6523. https://doi.org/10.1021/jp401243m.

(27) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. Determination of Electrostatic Parameters for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2005**, *1* (1), 153–168. https://doi.org/10.1021/ct049930p.

(28) Noskov, S. Yu.; Lamoureux, G.; Roux, B. Molecular Dynamics Study of Hydration in Ethanol−Water Mixtures Using a Polarizable Force Field. *J. Phys. Chem. B* **2005**, *109* (14), 6705–6713. https://doi.org/10.1021/jp045438q.

(29) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D. Polarizable Empirical Force Field for Alkanes Based on the Classical Drude Oscillator Model. *J. Phys. Chem. B* **2005**, *109* (40), 18988–18999. https://doi.org/10.1021/jp053182y.

(30) Anisimov, V. M.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. Polarizable Empirical Force Field for the Primary and Secondary Alcohol Series Based on the Classical Drude Model. *J. Chem. Theory Comput.* **2007**, *3* (6), 1927–1946. https://doi.org/10.1021/ct700100a.
(31) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D. Additive and Classical Drude Polarizable Force Fields for Linear and Cyclic Ethers. *J. Chem. Theory Comput.* **2007**, *3* (3), 1120–1133. https://doi.org/10.1021/ct600350s.
(32) Harder, E.; Anisimov, V. M.; Whitfield, T.; MacKerell, A. D.; Roux, B. Understanding the Dielectric Properties of Liquid Amides from a Polarizable Force Field. *J. Phys. Chem. B* **2008**, *112* (11), 3509–3521. https://doi.org/10.1021/jp709729d.
(33) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; Jr, A. D. M. Polarizable Empirical Force Field for Aromatic Compounds Based on the Classical Drude Oscillator. **2008**, 32.
(34) Lopes, P. E. M.; Lamoureux, G.; Mackerell, A. D. Polarizable Empirical Force Field for Nitrogen-Containing Heteroaromatic Compounds Based on the Classical Drude Oscillator. *J. Comput. Chem.* **2009**, *30* (12), 1821–1838. https://doi.org/10.1002/jcc.21183.
(35) Baker, C. M.; MacKerell, A. D. Polarizability Rescaling and Atom-Based Thole Scaling in the CHARMM Drude Polarizable Force Field for Ethers. *J. Mol. Model.* **2010**, *16* (3), 567–576. https://doi.org/10.1007/s00894-009-0572-4.
(36) Yu, H.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; MacKerell, A. D.; Roux, B. Simulating Monovalent and Divalent Ions in Aqueous Solution Using a Drude Polarizable Force Field. *J. Chem. Theory Comput.* **2010**, *6* (3), 774–786. https://doi.org/10.1021/ct900576a.
(37) Zhu, X.; Mackerell, A. D. Polarizable Empirical Force Field for Sulfur-Containing Compounds Based on the Classical Drude Oscillator Model. *J. Comput. Chem.* **2010**, NA-NA. https://doi.org/10.1002/jcc.21527.
(38) Chowdhary, J.; Harder, E.; Lopes, P. E. M.; Huang, L.; MacKerell, A. D.; Roux, B. A Polarizable Force Field of Dipalmitoylphosphatidylcholine Based on the Classical Drude Model for Molecular Dynamics Simulations of Lipids. *J. Phys. Chem. B* **2013**, *117* (31), 9142–9160. https://doi.org/10.1021/jp402860e.
(39) He, X.; Lopes, P. E. M.; MacKerell, A. D. Polarizable Empirical Force Field for Acyclic Polyalcohols Based on the Classical Drude Oscillator: Polarizable Force Field for Acyclic Polyols. *Biopolymers* **2013**, *99* (10), 724–738. https://doi.org/10.1002/bip.22286.
(40) Lin, B.; Lopes, P. E. M.; Roux, B.; MacKerell, A. D. Kirkwood-Buff Analysis of Aqueous *N* -Methylacetamide and Acetamide Solutions Modeled by the CHARMM Additive and Drude Polarizable Force Fields. *J. Chem. Phys.* **2013**, *139* (8), 084509. https://doi.org/10.1063/1.4818731.
(41) Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D. Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2013**, *9* (12), 5430–5449. https://doi.org/10.1021/ct400781b.
(42) Savelyev, A.; MacKerell, A. D. All-Atom Polarizable Force Field for DNA Based on the Classical Drude Oscillator Model. *J. Comput. Chem.* **2014**, *35* (16), 1219–1239. https://doi.org/10.1002/jcc.23611.
(43) Savelyev, A.; MacKerell, A. D. Balancing the Interactions of Ions, Water, and DNA in the Drude Polarizable Force Field. *J. Phys. Chem. B* **2014**, *118* (24), 6742–6757. https://doi.org/10.1021/jp503469s.

(44) Jana, M.; MacKerell, A. D. CHARMM Drude Polarizable Force Field for Aldopentofuranoses and Methyl-Aldopentofuranosides. *J. Phys. Chem. B* **2015**, *119* (25), 7846–7859. https://doi.org/10.1021/acs.jpcb.5b01767.
(45) Jana, M.; MacKerell, A. D. CHARMM Drude Polarizable Force Field for Aldopentofuranoses and Methyl-Aldopentofuranosides. *J. Phys. Chem. B* **2015**, *119* (25), 7846–7859. https://doi.org/10.1021/acs.jpcb.5b01767.
(46) Patel, D. S.; He, X.; MacKerell, A. D. Polarizable Empirical Force Field for Hexopyranose Monosaccharides Based on the Classical Drude Oscillator. *J. Phys. Chem. B* **2015**, *119* (3), 637–652. https://doi.org/10.1021/jp412696m.
(47) Lemkul, J. A.; MacKerell, A. D. Polarizable Force Field for DNA Based on the Classical Drude Oscillator: I. Refinement Using Quantum Mechanical Base Stacking and Conformational Energetics. *J. Chem. Theory Comput.* **2017**, *13* (5), 2053–2071. https://doi.org/10.1021/acs.jctc.7b00067.
(48) Lemkul, J. A.; MacKerell, A. D. Polarizable Force Field for DNA Based on the Classical Drude Oscillator: II. Microsecond Molecular Dynamics Simulations of Duplex DNA. *J. Chem. Theory Comput.* **2017**, *13* (5), 2072–2085. https://doi.org/10.1021/acs.jctc.7b00068.
(49) Li, H.; Chowdhary, J.; Huang, L.; He, X.; MacKerell, A. D.; Roux, B. Drude Polarizable Force Field for Molecular Dynamics Simulations of Saturated and Unsaturated Zwitterionic Lipids. *J. Chem. Theory Comput.* **2017**, *13* (9), 4535–4552. https://doi.org/10.1021/acs.jctc.7b00262.
(50) Small, M. C.; Aytenfisu, A. H.; Lin, F.-Y.; He, X.; MacKerell, A. D. Drude Polarizable Force Field for Aliphatic Ketones and Aldehydes, and Their Associated Acyclic Carbohydrates. *J. Comput. Aided Mol. Des.* **2017**, *31* (4), 349–363. https://doi.org/10.1007/s10822-017-0010-0.
(51) Lin, F.-Y.; Lopes, P. E. M.; Harder, E.; Roux, B.; MacKerell, A. D. Polarizable Force Field for Molecular Ions Based on the Classical Drude Oscillator. *J. Chem. Inf. Model.* **2018**, *58* (5), 993–1004. https://doi.org/10.1021/acs.jcim.8b00132.
(52) Lin, F.-Y.; MacKerell, A. D. Polarizable Empirical Force Field for Halogen-Containing Compounds Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2018**, *14* (2), 1083–1098. https://doi.org/10.1021/acs.jctc.7b01086.
(53) Yang, M.; Aytenfisu, A. H.; MacKerell, A. D. Proper Balance of Solvent-Solute and Solute-Solute Interactions in the Treatment of the Diffusion of Glucose Using the Drude Polarizable Force Field. *Carbohydr. Res.* **2018**, *457*, 41–50. https://doi.org/10.1016/j.carres.2018.01.004.
(54) Vanommeslaeghe, K.; Guvench, O.; MacKerell Jr, A. D. Molecular Mechanics. *Curr. Pharm. Des.* **2014**, *20* (20), 3281–3292.
(55) Mackerell, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25* (13), 1584–1604. https://doi.org/10.1002/jcc.20082.
(56) MacKerell, A. D.; Karplus, M. Importance of Attractive van Der Waals Contribution in Empirical Energy Function Models for the Heat of Vaporization of Polar Liquids. *J. Phys. Chem.* **1991**, *95* (26), 10559–10560. https://doi.org/10.1021/j100179a013.
(57) Gough, C. A.; Debolt, S. E.; Kollman, P. A. Derivation of Fluorine and Hydrogen Atom Parameters Using Liquid Simulations. *J. Comput. Chem.* **1992**, *13* (8), 963–970. https://doi.org/10.1002/jcc.540130806.
(58) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236. https://doi.org/10.1021/ja9621760.

(59) Fox, T.; Kollman, P. A. Application of the RESP Methodology in the Parametrization of Organic Solvents. *J. Phys. Chem. B* **1998**, *102* (41), 8070–8079. https://doi.org/10.1021/jp9717655.
(60) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. Free Energies of Hydration and Pure Liquid Properties of Hydrocarbons from the OPLS All-Atom Model. *J. Phys. Chem.* **1994**, *98* (49), 13077–13082. https://doi.org/10.1021/j100100a043.
(61) Rizzo, R. C.; Jorgensen, W. L. OPLS All-Atom Model for Amines: Resolution of the Amine Hydration Problem. *J. Am. Chem. Soc.* **1999**, *121* (20), 4827–4836. https://doi.org/10.1021/ja984106u.
(62) Chen, I. J.; Yin, D.; MacKerell, A. D. Combined *Ab initio* Empirical Approach for Optimization of Lennard-Jones Parameters for Polar-Neutral Compounds. *J. Comput. Chem.* **2002**, *23* (2), 199–213. https://doi.org/10.1002/jcc.1166.
(63) Yin, D.; MacKerell Jr., A. D. Combined *Ab initio*/Empirical Approach for Optimization of Lennard–Jones Parameters. *J. Comput. Chem.* **1998**, *19* (3), 334–348. https://doi.org/10.1002/(SICI)1096-987X(199802)19:3<334::AID-JCC7>3.0.CO;2-U.
(64) Levitt, M.; Lifson, S. Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure. *J. Mol. Biol.* **1969**, *46* (2), 269–279. https://doi.org/10.1016/0022-2836(69)90421-5.
(65) Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, *58* (3), 565–578. https://doi.org/10.1021/acs.jcim.8b00042.
(66) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197. https://doi.org/10.1021/ja00124a002.
(67) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. Ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16* (1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591.
(68) Robertson, M. J.; Qian, Y.; Robinson, M. C.; Tirado-Rives, J.; Jorgensen, W. L. Development and Testing of the OPLS-AA/M Force Field for RNA. *J. Chem. Theory Comput.* **2019**, *15* (4), 2734–2742. https://doi.org/10.1021/acs.jctc.9b00054.
(69) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell Jr, 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. https://doi.org/10.1021/ct300400x.
(70) 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. https://doi.org/10.1021/jp973084f.
(71) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D. CHARMM36m: An Improved Force Field

for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14* (1), 71–73. https://doi.org/10.1038/nmeth.4067.

(72) 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. https://doi.org/10.1007/s10822-018-0134-x.

(73) Roux, B.; Simonson, T. Implicit Solvent Models. *Biophys. Chem.* **1999**, *78* (1), 1–20. https://doi.org/10.1016/S0301-4622(98)00226-9.

(74) Dauber-Osguthorpe, P.; Hagler, A. T. Biomolecular Force Fields: Where Have We Been, Where Are We Now, Where Do We Need to Go and How Do We Get There? *J. Comput. Aided Mol. Des.* **2019**, *33* (2), 133–203. https://doi.org/10.1007/s10822-018-0111-4.

(75) Bernal, J. D.; Fowler, R. H. A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. *J. Chem. Phys.* **1933**, *1* (8), 515–548. https://doi.org/10.1063/1.1749327.

(76) Rowlinson, J. S. The Lattice Energy of Ice and the Second Virial Coefficient of Water Vapour. *Trans. Faraday Soc.* **1951**, *47* (0), 120–129. https://doi.org/10.1039/TF9514700120.

(77) Ben-Naim, A.; Stihinger, F. H.; Hill, M. Aspects Ofthe Statistical-Mechanical Theory of Water. 36.

(78) Jorgensen, W. L. Quantum and Statistical Mechanical Studies of Liquids. 10. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103* (2), 335–340. https://doi.org/10.1021/ja00392a016.

(79) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981*; Pullman, B., Ed.; The Jerusalem Symposia on Quantum Chemistry and Biochemistry; Springer Netherlands: Dordrecht, 1981; pp 331–342. https://doi.org/10.1007/978-94-015-7658-1_21.

(80) 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. https://doi.org/10.1063/1.445869.

(81) Stillinger, F. H.; Rahman, A. Improved Simulation of Liquid Water by Molecular Dynamics. *J. Chem. Phys.* **1974**, *60* (4), 1545–1557. https://doi.org/10.1063/1.1681229.

(82) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217. https://doi.org/10.1002/jcc.540040211.

(83) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems. *J. Chem. Phys.* **1996**, *105* (5), 1902–1921. https://doi.org/10.1063/1.472061.

(84) MacKerell, A. D.; Wiorkiewicz-Kuczera, J.; Karplus, M. An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids. *J. Am. Chem. Soc.* **1995**, *117* (48), 11946–11975. https://doi.org/10.1021/ja00153a017.

(85) Schlenkrich, M.; Brickmann, J.; MacKerell, A. D.; Karplus, M. An Empirical Potential Energy Function for Phospholipids: Criteria for Parameter Optimization and Applications. In *Biological Membranes: A Molecular Perspective from Computation and Experiment*; Merz, K. M., Roux, B., Eds.;

Birkhäuser: Boston, MA, 1996; pp 31–81. https://doi.org/10.1007/978-1-4684-8580-6_2.

(86) Mackerell Jr., A. D.; Feig, M.; Brooks III, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25* (11), 1400–1415. https://doi.org/10.1002/jcc.20065.

(87) MacKerell Jr, 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. https://doi.org/10.1021/ja036959e.

(88) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(89) Foloppe, N.; MacKerell Jr, A. D. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, *21* (2), 86–104. https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2<86::AID-JCC2>3.0.CO;2-G.

(90) MacKerell Jr., A. D.; Banavali, N. K. All-Atom Empirical Force Field for Nucleic Acids: II. Application to Molecular Dynamics Simulations of DNA and RNA in Solution. *J. Comput. Chem.* **2000**, *21* (2), 105–120. https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2<105::AID-JCC3>3.0.CO;2-P.

(91) Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J. Chem. Theory Comput.* **2009**, *5* (9), 2353–2370. https://doi.org/10.1021/ct900242e.

(92) *Biomolecular Simulations: Methods and Protocols*; Bonomi, M., Camilloni, C., Eds.; Methods in Molecular Biology; Springer New York: New York, NY, 2019; Vol. 2022. https://doi.org/10.1007/978-1-4939-9608-7.

(93) Felder, C. E.; Prilusky, J.; Silman, I.; Sussman, J. L. A Server and Database for Dipole Moments of Proteins. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W512–W521. https://doi.org/10.1093/nar/gkm307.

(94) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(95) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell Jr, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31* (4), 671–690. https://doi.org/10.1002/jcc.21367.

(96) Qiu, Y.; Smith, D.; Boothroyd, S.; Jang, H.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C.; Rizzi, A.; Lucas, X.; Tjanaka, B.; Shirts, M. R.; Gilson, M.; Chodera, J.; Bayly, C. I.; Mobley, D.; Wang, L.-P. Development and Benchmarking of Open Force Field v1.0.0, the Parsley Small Molecule Force Field. **2021**. https://doi.org/10.26434/chemrxiv-2021-l070l-v4.

(97) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12* (1), 281–296. https://doi.org/10.1021/acs.jctc.5b00864.

(98) Wang, J.; Wang, W.; Kollman, P. A.; Wang, C. T. J. 1 Submitted to Journal of Chemical Information and Computer Sciences Antechamber, An Accessory Software Package For Molecular Mechanical Calculations.
(99) 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. https://doi.org/10.1021/ci300363c.
(100) 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. https://doi.org/10.1021/ci3003649.
(101) 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. https://doi.org/10.1186/1758-2946-3-33.
(102) Croitoru, A.; Park, S.-J.; Kumar, A.; Lee, J.; Im, W.; MacKerell Jr, A. D.; Aleksandrov, A. Additive CHARMM36 Force Field for Nonstandard Amino Acids. *J. Chem. Theory Comput.* **2021**, *17* (6), 3554–3570. https://doi.org/10.1021/acs.jctc.1c00254.
(103) Croitoru, A.; Babin, M.; Myllykallio, H.; Gondry, M.; Aleksandrov, A. Cyclodipeptide Synthases of the NYH Subfamily Recognize TRNA Using an α-Helix Enriched with Positive Residues. *Biochemistry* **2021**, *60* (1), 64–76. https://doi.org/10.1021/acs.biochem.0c00761.
(104) 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. https://doi.org/10.1016/j.bbagen.2014.08.004.
(105) CGENff-Server.
(106) 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; D. J. Fox. Gaussian, 2009.
(107) Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.0.
(108) 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. https://doi.org/10.1080/00268977000101561.
(109) Xu, Y.; Vanommeslaeghe, K.; Aleksandrov, A.; MacKerell Jr, A. D.; Nilsson, L. Additive CHARMM Force Field for Naturally Occurring Modified Ribonucleotides. *J. Comput. Chem.* **2016**, *37* (10), 896–912. https://doi.org/10.1002/jcc.24307.
(110) 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. https://doi.org/10.1021/ct4003477.

(111) 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.
(112) Brooks, B. R.; Brooks, C. L.; MacKerell Jr, 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. https://doi.org/10.1002/jcc.21287.
(113) Vanommeslaeghe, K.; Yang, M.; MacKerell Jr, A. D. Robustness in the Fitting of Molecular Mechanics Parameters. *J. Comput. Chem.* **2015**, *36* (14), 1083–1101. https://doi.org/10.1002/jcc.23897.
(114) Aleksandrov, A. A Molecular Mechanics Model for Flavins. *J. Comput. Chem.* **2019**, *40* (32), 2834–2842. https://doi.org/10.1002/jcc.26061.
(115) 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. https://doi.org/10.1021/ct100578z.
(116) 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* (Web Server issue), W665-667.
(117) 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. https://doi.org/10.1002/jcc.20289.
(118) Darden, T. Treatment of Long-Range Forces and Potential. In *Computational biochemistry and biophysics*; Marcel Dekker: New York, NY, 2001.
(119) 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. https://doi.org/10.1063/1.448118.
(120) 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. https://doi.org/10.1016/j.jmb.2008.03.022.
(121) 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. https://doi.org/10.1021/ja0741933.
(122) Gfeller, D.; Michielin, O.; Zoete, V. SwissSidechain: A Molecular and Structural Database of Non-Natural Sidechains. *Nucleic Acids Res.* **2013**, *41* (D1), D327–D332. https://doi.org/10.1093/nar/gks991.
(123) 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. https://doi.org/10.1093/bioinformatics/btu789.
(124) 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. https://doi.org/10.1038/nature10154.

(125) Tapiero, H.; Townsend, D. M.; Tew, K. D. The Antioxidant Role of Selenium and Seleno-Compounds. *Biomed. Pharmacother. Biomedecine Pharmacother.* **2003**, *57* (3–4), 134–144.
(126) Tinggi, U. Selenium: Its Role as Antioxidant in Human Health. *Environ. Health Prev. Med.* **2008**, *13* (2), 102–108. https://doi.org/10.1007/s12199-007-0019-4.
(127) Hendrickson, W. A. Maturation of MAD Phasing for the Determination of Macromolecular Structures. *J. Synchrotron Radiat.* **1999**, *6* (4), 845–851. https://doi.org/10.1107/S0909049599007591.
(128) 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. https://doi.org/10.1021/bi0479205.
(129) MarvinSketch, 2019.
(130) 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* (bau116). https://doi.org/10.1093/database/bau116.
(131) Gondry, M.; Sauguet, L.; Belin, P.; Thai, R.; Amouroux, R.; Tellier, C.; Tuphile, K.; Jacquet, M.; Braud, S.; Courçon, M.; Masson, C.; Dubois, S.; Lautru, S.; Lecoq, A.; Hashimoto, S.; Genet, R.; Pernodet, J.-L. Cyclodipeptide Synthases Are a Family of TRNA-Dependent Peptide Bond–Forming Enzymes. *Nat. Chem. Biol.* **2009**, *5* (6), 414–420. https://doi.org/10.1038/nchembio.175.
(132) Moutiez, M.; Belin, P.; Gondry, M. Aminoacyl-TRNA-Utilizing Enzymes in Natural Product Biosynthesis. *Chem. Rev.* **2017**, *117* (8), 5578–5618. https://doi.org/10.1021/acs.chemrev.6b00523.
(133) Bonnefond, L.; Arai, T.; Sakaguchi, Y.; Suzuki, T.; Ishitani, R.; Nureki, O. Structural Basis for Nonribosomal Peptide Synthesis by an Aminoacyl-TRNA Synthetase Paralog. *Proc. Natl. Acad. Sci.* **2011**, *108* (10), 3912–3917. https://doi.org/10.1073/pnas.1019480108.
(134) Bourgeois, G.; Seguin, J.; Babin, M.; Belin, P.; Moutiez, M.; Mechulam, Y.; Gondry, M.; Schmitt, E. Structural Basis for Partition of the Cyclodipeptide Synthases into Two Subfamilies. *J. Struct. Biol.* **2018**, *203* (1), 17–26. https://doi.org/10.1016/j.jsb.2018.03.001.
(135) Sauguet, L.; Moutiez, M.; Li, Y.; Belin, P.; Seguin, J.; Le Du, M.-H.; Thai, R.; Masson, C.; Fonvielle, M.; Pernodet, J.-L.; Charbonnier, J.-B.; Gondry, M. Cyclodipeptide Synthases, a Family of Class-I Aminoacyl-TRNA Synthetase-like Enzymes Involved in Non-Ribosomal Peptide Synthesis. *Nucleic Acids Res.* **2011**, *39* (10), 4475–4489. https://doi.org/10.1093/nar/gkr027.
(136) Vetting, M. W.; Hegde, S. S.; Blanchard, J. S. The Structure and Mechanism of the Mycobacterium Tuberculosis Cyclodityrosine Synthetase. *Nat. Chem. Biol.* **2010**, *6* (11), 797–799. https://doi.org/10.1038/nchembio.440.
(137) Moutiez, M.; Schmitt, E.; Seguin, J.; Thai, R.; Favry, E.; Belin, P.; Mechulam, Y.; Gondry, M. Unravelling the Mechanism of Non-Ribosomal Peptide Synthesis by Cyclodipeptide Synthases. *Nat. Commun.* **2014**, *5* (1), 5141. https://doi.org/10.1038/ncomms6141.
(138) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS4: Improving Force Field Accuracy on Challenging Regimes of

Chemical Space. *J. Chem. Theory Comput.* **2021**, *17* (7), 4291–4300. https://doi.org/10.1021/acs.jctc.1c00302.

(139) Seminario, J. M. Calculation of Intramolecular Force Fields from Second-Derivative Tensors. *Int. J. Quantum Chem.* **1996**, *60* (7), 1271–1277. https://doi.org/10.1002/(SICI)1097-461X(1996)60:7<1271::AID-QUA8>3.0.CO;2-W.

(140) Wang, L.-P.; Chen, J.; Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *J. Chem. Theory Comput.* **2013**, *9* (1), 452–460. https://doi.org/10.1021/ct300826t.

(141) Hudson, P. S.; Boresch, S.; Rogers, D. M.; Woodcock, H. L. Accelerating QM/MM Free Energy Computations via Intramolecular Force Matching. *J. Chem. Theory Comput.* **2018**, *14* (12), 6327–6335. https://doi.org/10.1021/acs.jctc.8b00517.

(142) Hudson, P. S.; Han, K.; Woodcock, H. L.; Brooks, B. R. Force Matching as a Stepping Stone to QM/MM CB[8] Host/Guest Binding Free Energies: A SAMPL6 Cautionary Tale. *J. Comput. Aided Mol. Des.* **2018**, *32* (10), 983–999. https://doi.org/10.1007/s10822-018-0165-3.

(143) Akin-Ojo, O.; Song, Y.; Wang, F. Developing *Ab initio* Quality Force Fields from Condensed Phase Quantum-Mechanics/Molecular-Mechanics Calculations through the Adaptive Force Matching Method. *J. Chem. Phys.* **2008**, *129* (6), 064108. https://doi.org/10.1063/1.2965882.

(144) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S. Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J. Phys. Chem. B* **2017**, *121* (16), 4023–4039. https://doi.org/10.1021/acs.jpcb.7b02320.

(145) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. Systematic *Ab initio* Gradient Calculation of Molecular Geometries, Force Constants, and Dipole Moment Derivatives. *J. Am. Chem. Soc.* **1979**, *101* (10), 2550–2560. https://doi.org/10.1021/ja00504a009.

(146) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(147) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *J. Comput. Chem.* **2013**, *34* (32), 2757–2770. https://doi.org/10.1002/jcc.23422.

(148) Hopkins, C. W.; Roitberg, A. E. Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem. *J. Chem. Inf. Model.* **2014**, *54* (7), 1978–1986. https://doi.org/10.1021/ci500112w.

(149) Urwin, D. J.; Alexandrova, A. N. Regularization of Least Squares Problems in CHARMM Parameter Optimization by Truncated Singular Value Decompositions. *J. Chem. Phys.* **2021**, *154* (18), 184101. https://doi.org/10.1063/5.0045982.

(150) Bacskay, G. B. A Quadratically Convergent Hartree—Fock (QC-SCF) Method. Application to Closed Shell Systems. *Chem. Phys.* **1981**, *61* (3), 385–404. https://doi.org/10.1016/0301-0104(81)85156-7.

(151) Johnson, R. D. *NIST Computational Chemistry Comparison and Benchmark Database*. NIST Computational Chemistry Comparison and Benchmark Database. http://cccbdb.nist.gov/ (accessed 2022-01-15).

(152) Allen, A. E. A.; Payne, M. C.; Cole, D. J. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.* **2018**, *14* (1), 274–281. https://doi.org/10.1021/acs.jctc.7b00785.

(153) *Compound Sourcing, Selling and Purchasing Platform*. Molport. https://www.molport.com/shop/index?version=2 (accessed 2022-07-20).
(154) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675.
(155) *Home - Enamine*. https://enamine.net/ (accessed 2022-07-20).
(156) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and Searching 10(20) Synthetically Accessible Structures. *J. Comput. Aided Mol. Des.* **2007**, *21* (6), 341–350. https://doi.org/10.1007/s10822-006-9093-8.
(157) Chevillard, F.; Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **2015**, *55* (9), 1824–1835. https://doi.org/10.1021/acs.jcim.5b00203.
(158) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.-A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. HMDB: The Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35* (Database issue), D521-526. https://doi.org/10.1093/nar/gkl923.
(159) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
(160) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys. PCCP* **2010**, *12* (40), 12899–12908. https://doi.org/10.1039/c0cp00151a.
(161) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* **2015**, *10* (5), 449–461. https://doi.org/10.1517/17460441.2015.1032936.
(162) Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20* (7), 13384–13421. https://doi.org/10.3390/molecules200713384.
(163) Goel, H.; Hazel, A.; Yu, W.; Jo, S.; MacKerell, A. D. Application of Site-Identification by Ligand Competitive Saturation in Computer-Aided Drug Design. *New J. Chem. Nouv. J. Chim.* **2022**, *46* (3), 919–932. https://doi.org/10.1039/d1nj04028f.
(164) van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59* (2), 644–651. https://doi.org/10.1021/acs.jcim.8b00737.
(165) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875. https://doi.org/10.1021/ci300415d.

(166) *CAS REGISTRY*. CAS. https://www.cas.org/cas-data/cas-registry (accessed 2022-07-20).
(167) Reymond, J.-L. The Chemical Space Project. *Acc Chem Res* **2015**, 9.
(168) Humbeck, L.; Weigang, S.; Schäfer, T.; Mutzel, P.; Koch, O. CHIPMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein-Protein Interaction Modulators. *ChemMedChem* **2018**, *13* (6), 532–539. https://doi.org/10.1002/cmdc.201700689.
(169) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2. ed.; Wiley: Chichester, 2014.
(170) Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, *56* (7), 1253–1266. https://doi.org/10.1021/acs.jcim.6b00173.
(171) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182. https://doi.org/10.1021/ci049714.
(172) *ChEMBL Database*. https://www.ebi.ac.uk/chembl/ (accessed 2022-07-20).
(173) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34* (Database issue), D668-672. https://doi.org/10.1093/nar/gkj067.
(174) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.
(175) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; Mclay, I. M.; Bradshaw, J. Drug Rings Database with Web Interface. A Tool for Identifying Alternative Chemical Rings in Lead Discovery Programs. *J. Med. Chem.* **2003**, *46* (15), 3257–3274. https://doi.org/10.1021/jm0300429.
(176) Shearer, J.; Castro, J. L.; Lawson, A. D. G.; MacCoss, M.; Taylor, R. D. Rings in Clinical Trials and Drugs: Present and Future. *J. Med. Chem.* **2022**, acs.jmedchem.2c00473. https://doi.org/10.1021/acs.jmedchem.2c00473.
(177) *RDKit*. http://www.rdkit.org/ (accessed 2022-07-20).
(178) *Sulstice/global-chem: A Chemical Knowledge Graph of What is Common in the World*. https://github.com/Sulstice/global-chem (accessed 2022-07-20).
(179) Fulmer, G. R.; Miller, A. J. M.; Sherden, N. H.; Gottlieb, H. E.; Nudelman, A.; Stoltz, B. M.; Bercaw, J. E.; Goldberg, K. I. NMR Chemical Shifts of Trace Impurities: Common Laboratory Solvents, Organics, and Gases in Deuterated Solvents Relevant to the Organometallic Chemist. *Organometallics* **2010**, *29* (9), 2176–2179. https://doi.org/10.1021/om100106e.
(180) *OpenSMILES Home Page*. http://opensmiles.org/ (accessed 2022-07-20).
(181) Hiorns, R. C.; Boucher, R. J.; Duhlev, R.; Hellwich, K.-H.; Hodge, P.; Jenkins, A. D.; Jones, R. G.; Kahovec, J.; Moad, G.; Ober, C. K.; Smith, D. W.; Stepto, R. F. T.; Vairon, J.-P.; Vohlídal, J. A brief guide to polymer nomenclature (IUPAC Technical Report). *Pure Appl. Chem.* **2012**, *84* (10), 2167–2169. https://doi.org/10.1351/PAC-REP-12-03-05.
(182) Broughton, H. B.; Watson, I. A. Selection of Heterocycles for Drug Design. *J. Mol. Graph. Model.* **2004**, *23* (1), 51–58. https://doi.org/10.1016/j.jmgm.2004.03.016.
(183) Taylor, R. D.; MacCoss, M.; Lawson, A. D. G. Rings in Drugs: Miniperspective. *J. Med. Chem.* **2014**, *57* (14), 5845–5859. https://doi.org/10.1021/jm4017625.
(184) Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. Privileged Scaffolds for Library Design and Drug Discovery. *Curr. Opin. Chem. Biol.* **2010**, *14* (3), 347–361. https://doi.org/10.1016/j.cbpa.2010.02.018.

(185) Gehringer, M.; Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2019**, *62* (12), 5673–5724. https://doi.org/10.1021/acs.jmedchem.8b01153.

(186) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72* (2), 171–179. https://doi.org/10.1107/S2052520616003954.

ECOLE DOCTORALE

Titre : Développement de méthodes de champ de force

Mots clés : champ de force ; CHARMM ; ZINC20 ; acide aminés non standards

Résumé : La découverte et le développement de molécules actives sont des processus coûteux en termes de ressources et temps qui peuvent être améliorés par la conception de médicaments assistée par ordinateur. Le but de cette thèse a été de développer le champ de force pour un large nombre de molécules d'importance biologique.

Dans un premier temps, je me suis concentrée sur l'extension du champ de force CHARMM à un large ensemble de 333 acides aminés non standards. J'ai considéré des acides aminés avec des chaînes latérales ainsi que des acides aminés avec les squelettes carbonés modifiés. Les termes inter- et intramoléculaires ont été paramétrés en ciblant des données *ab initio*. Une attention particulière a été donnée aux angles dièdres. La validation a été effectuée par des simulations de dynamique moléculaire de 20 systèmes protéiques.

Dans la deuxième partie, j'ai testé la mise en œuvre d'une nouvelle méthode d'optimisation des termes de liaison et d'angle de valence. Pour améliorer la transférabilité et la robustesse des paramètres développés, les déviations structurelles entre les structures *ab initio* et CHARMM ont été autorisées pendant l'optimisation.

Dans la dernière partie du projet, j'ai effectué une paramétrisation à grande échelle de ~300 000 ligands de la base de données ZINC20. En utilisant le tri basé sur les termes de liaison, la taille de la bibliothèque a été réduite à 7 387 molécules. Une attention particulière a été accordée à l'optimisation des cycles non planaires.

Dans l'ensemble, cette thèse constitue une immense extension du champs de force CHARMM et sera d'un grand intérêt pour la communité scientifique.

Title : Development of Force Field Methods

Keywords : force field ; CHARMM ; ZINC20 ; nonstandard amino acids

Abstract : Drug discovery and development are very time and resources consuming processes, which can be significantly facilitated by force field-based computer-aided drug design. The goal of this PhD was to develop the force field for a large number of biologically important molecules.

In the first part, I focused on extending the CHARMM force field to a large set of 333 nonstandard amino acids. I included amino acids with nonstandard side chains as well as amino acids with modified backbone groups. Both inter- and intramolecular terms were parametrized targeting *ab initio* data. A special emphasis was given to rotatable dihedrals. Validation was performed by molecular dynamics simulations of 20 protein systems.

In the second part, I tested the implementation of a new method for bond and valence angle terms optimization. To improve transferability and robustness of developed parameters, structural deviations between *ab initio* and CHARMM structures were allowed during optimization.

In the final part of the project, I performed large-scale parametrization of ~300 000 ligands from ZINC20 database. Based on bond term sorting the size of the library was decreased to 7 387 molecules. A special attention was given to the optimization of non-planar rings.

Overall, this work represents an immense extension for CHARMM force field and will be of great use for the scientific community.

Institut Polytechnique de Paris
91120 Palaiseau, France