JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Polarizable Atomic Multipole X-Ray Refinement: Particle Mesh Ewald Electrostatics for Macromolecular Crystals

Michael J. Schnieders,*[,†] Timothy D. Fenn,[‡,§] and Vijay S. Pande[†]

[†]Department of Chemistry

[‡]Department of Molecular and Cellular Physiology

Stanford University, Stanford California 94305, United States

[§]Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789, United States

Supporting Information

ABSTRACT: Refinement of macromolecular models from X-ray crystallography experiments benefits from prior chemical knowledge at all resolutions. As the quality of the prior chemical knowledge from quantum or classical molecular physics improves, in principle so will resulting structural models. Due to limitations in computer performance and electrostatic algorithms, commonly used macromolecules X-ray crystallography refinement protocols have had limited support for rigorous molecular physics in the past. For example, electrostatics is often neglected in favor of nonbonded interactions based on a purely repulsive van der Waals potential. In this work we present advanced algorithms for desktop workstations that open the door to X-ray refinement of even the most challenging macromolecular data sets using state-of-the-art classical molecular physics. First we describe theory for particle mesh Ewald (PME) summation that consistently handles the symmetry of all 230 space groups, replicates of the unit cell such that the minimum image convention can be used with a real space cutoff of any size and the combination of space group symmetry with replicates. An implementation of symmetry accelerated PME for the polarizable atomic multipole optimized energetics for biomolecular applications (AMOEBA) force field is presented. Relative to a single CPU core performing calculations on a P1 unit cell, our AMOEBA engine called Force Field X (FFX) accelerates energy evaluations by more than a factor of 24 on an 8-core workstation with a Tesla GPU coprocessor for 30 structures that contain 240 000 atoms on average in the unit cell. The benefit of AMOEBA electrostatics evaluated with PME for macromolecular X-ray crystallography refinement is demonstrated via rerefinement of 10 crystallographic data sets that range in resolution from 1.7 to 4.5 Å. Beginning from structures obtained by local optimization without electrostatics, further optimization using AMOEBA with PME electrostatics improved agreement of the model with the data (R_{free} was lowered by 0.5%), improved geometric features such as favorable (ϕ, ψ) backbone conformations, and lowered the average potential energy per residue by over 10 kcal/mol. Furthermore, the MolProbity structure validation tool indicates that the geometry of these rerefined structures is consistent with X-ray crystallographic data collected up to 2.2 Å, which is 0.9 Å better than the actual mean quality (3.1 Å). We conclude that polarizable AMOEBA-assisted X-ray refinement offers advantages to methods that neglect electrostatics and is now efficient enough for routine use.

I. INTRODUCTION

We recently described theory for biomolecular X-ray crystallography refinement based on optimization of a target function E_{target} that is the sum of polarizable atomic multipole descriptions of both the X-ray scattering $E_{X\text{-}ray}$ and the chemical potential energy $E_{chemistry}$

$$E_{target} = w_a E_{X\text{-}ray} + E_{chemistry} \quad (1)$$

where the weight w_a controls their relative importance.[1] The method has been successfully applied to ultrahigh-resolution (0.5 Å) peptide crystals[1] and high-resolution (1.0 Å) protein and nucleic acid biomolecules[2] and to neutron crystallography.[3] In the first two cases, our Cartesian Gaussian multipolar scattering model with multipole coefficients from the atomic multipole optimized energetics for biomolecular applications (AMOEBA) force field[4−8] improved R/R_{free} statistics.[1,2] For the biomolecular and neutron crystallography data sets, the polarizable AMOEBA energetic model was shown to be critical for refinement of water hydrogen-bonding networks.[2,3] These encouraging results motivate further work to apply AMOEBA-assisted X-ray refinement to larger macromolecular crystals at lower resolution (1.0−4.5 Å). However, our rough draft implementation based on the combination of TINKER v. 5.0[9] and CNS v. 1.21[10] required expansion to P1 for AMOEBA forces and was limited to systems with on the order of 25 000 atoms. To address this, the present work describes a completely new implementation of AMOEBA ($E_{chemistry}$) designed from the ground up for application to large biomolecular data sets on modern desktop workstations. We describe particle mesh Ewald (PME) electrostatics theory that consistently supports all 230 space groups, replicates of the unit cell such that the minimum image convention can be used with a real space cutoff of any size and the combination of space group symmetry with replicates.[11] Our implementation, called Force Field X (FFX), uses the Java Runtime Environment (JRE) for

Received: September 4, 2010
Published: March 09, 2011

shared memory parallelization across CPU cores in combination with offloading computational work to a GPU coprocessor.

A rigorous solution for the electrostatic potential within an infinite lattice of unit cells was originally described by Ewald in 1921.[12] The method, now referred to as Ewald summation, converts the real space Coulomb lattice summation into the sum of real space and reciprocal space contributions. PME, introduced by Darden et al. in 1993,[11] formulated the reciprocal space portion using Lagrange interpolation and fast Fourier transforms (FFT) to achieve $N\cdot\log(N)$ scaling for the calculation of structure factors. Smooth PME[13] replaced Lagrange interpolation with B-spline interpolation, which offers analytic gradients of arbitrary accuracy and extension to multipolar charge descriptions.[14−16] Recently, Gaussian split Ewald,[17] multilevel Ewald,[18] and interlaced[19,20] approaches have been presented to improve the scaling and/or parallelization of particle mesh methods.

Computation of structure factors is of fundamental importance to both PME electrostatics and X-ray scattering. In both cases this is a direct consequence of the tiling of three-dimensional (3D) space by a repeating unit cell. In the context of X-ray refinement, structure factors computed from the structural model are formally compared to those measured experimentally within the X-ray term ($E_{\text{X-ray}}$) of the overall refinement target given in eq 1. Earlier work based on the FFX platform presented a differentiable X-ray term that includes the scattering of bulk solvent, which is used for the AMOEBA-assisted rerefinements presented later.[21] Within the context of force field potentials ($E_{\text{chemistry}}$), periodic boundary conditions (PBC) facilitate the study of an infinitely large system and eliminate edge effects inherent to aperiodic descriptions.

Use of PME for molecular dynamics simulations has flourished because it is often the most efficient way to avoid artifacts introduced by truncation schemes.[14,22] The importance of electrostatics to biomolecular energetics has led to the development of schemes to decompose electron density[23,24] into relatively many low-order sites (i.e., charges at atomic centers, bond centers, lone pairs, etc.) vs fewer higher order atomic multipole sites.[25−29] Generally, there has been greater emphasis placed on modeling lone pair electron density than on bonding electron density in order to predict hydrogen bonding.[5] From a crystallography perspective, bond density is of greater consequence since it contributes more to X-ray scattering than lone pair density for the majority of biomolecular structures.[30] For example, the phenol ring of tyrosine has seven bonds between heavy atoms but only a single lone pair site. To place electron density at bond centers for tetrahedral chemistries, diamond for example, either an atomic multipole expansion through hexadecapole order or bond charges is necessary.[1,31]

On the other hand, use of Ewald summation for X-ray crystallography refinement has lagged behind its adoption for molecular simulation. A first step toward including electrostatics within refinement by simulated annealing was described by Weis et al. for influenza virus hemagglutinin; however, the lattice summation was evaluated using a conditionally convergent spherical cutoff.[32] This approach was reintroduced[33] with the addition of an analytic generalized Born continuum solvent,[34,35] however, the underlying conditional convergence of the Coulomb lattice summation was not addressed. Although fixed charge force fields may be designed for use with a spherical cutoff, this approach is based on the observation that the radial distribution function (RDF) of neat organic liquids asymptote to unity at about one nanometer.[36] However, the RDFs for molecules within a periodic crystal do not decay to one but have periodic peaks at all lengths scales. Long-range correlations must be considered or inclusion of electrostatics can lead to systematic errors.[14,22] Currently, refinement packages, such as CNS,[10,37] Phenix,[38,39] BUSTER,[40] and Refmac[41] lack a rigorous Coulomb lattice summation method. Therefore, a key motivation for the current work was to create a state-of-the-art force field engine that can be incorporated into existing X-ray crystallography software and can handle data sets of any size—from small molecule crystals to ribosome crystals with millions of atoms.

We begin by describing the explicit incorporation of symmetry operators into PME electrostatics for the AMOEBA force field.[8] Our algorithm consistently accommodates not only space groups but also replicates of the central unit cell and the combination of space group symmetry with replicates. Details of our parallelization scheme are presented for shared memory parallelization over CPU cores in combination with the option to offload the PME reciprocal space sum to a GPU coprocessor. Overall timings and the speed up relative to expansion to P1 for 30 crystals with a variety of space groups are discussed in order to demonstrate that for the first time PME electrostatics are affordable for X-ray refinement of all system sizes encountered in macromolecular crystallography. Finally, we compare rerefinement of 10 X-ray crystallography data sets with and without polarizable AMOEBA electrostatics.

II. PARTICLE MESH EWALD WITH SPACE GROUP SYMMETRY

A. Unit Cell, Space Group, and Asymmetric Unit Definitions. We define a lattice Λ in direct space by its basis vectors **a**, **b**, and **c** that have Euclidean lengths a, b, and c, respectively. The Cartesian component of a vector will be denoted by a subscript, for example a_α, where $\alpha \in \{1,2,3\}$. The conjugate reciprocal lattice Λ^* is defined by basis vectors $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$ that have Euclidean lengths a^*, b^*, and c^*, respectively. The unit cell U of the lattice Λ consists of all points $\mathbf{r}_{\text{frac}}$ that have fractional coordinates with $0 \le r_{\text{frac},\alpha} \le 1$. Cartesian coordinates **r** can be converted to fractional coordinates $\mathbf{r}_{\text{frac}}$ via multiplication by a 3×3 fractionalization matrix $\mathbf{r}_{\text{frac}} = \mathbf{r}^t\mathbf{A}$, where the superscript t denotes the transpose of the Cartesian coordinate column vector into a row vector and the columns of **A** are the reciprocal basis vectors. The inverse operation is given by $\mathbf{r} = \mathbf{r}_{\text{frac}}^t\mathbf{A}^{-1}$, where the rows of $\mathbf{A}^{-1}$ are given by the direct basis vectors.

The space group of a crystal will be defined by its set of n_s symmetry operators. The i^{th} fractional symmetry operator $(\boldsymbol{R}_i,\boldsymbol{t}_i)$ is composed of a 3×3 rotation matrix $\mathbf{R}_i$ plus a translation vector $\mathbf{t}_i$. The analogous Cartesian symmetry operators will be denoted in *italics* as $(\boldsymbol{R}_i,\boldsymbol{t}_i)$. We assume a chemical system that is modeled by a set of n_a unique atoms, which constitute the asymmetric unit. The position of each atom i is described by orthogonal coordinates $\mathbf{r}_i$ and its permanent atomic charge, dipole, and quadrupole by $\{q_i,\mathbf{d}_i,\Phi_i\}$. The coordinates of any atom in the unit cell can then be generated by application of a symmetry operator to one of the atoms in the unique set. The atomic electrostatic moments also require application of the rotational part of the symmetry operator. To avoid unnecessary complexity and to keep the presentation as general as possible, discussion of the AMOEBA self-consistent field procedure that generates an induced dipole $\mathbf{u}_i$ at each multipole site is restricted to the Supporting Information.

B. Replicates of the Unit Cell. Application of the minimum image convention to a unit cell whose smallest width is less than half the real space cutoff r_{cut} requires generation of atomic coordinates in replicates of the unit cell.[42] The concept of symmetry operators can be generalized to a replicated super cell with $n_u = m_1 \times m_2 \times m_3$ copies of the unit cell arranged along scaled direct space basis vectors $\{\mathbf{a}_r = \mathbf{a}m_1, \mathbf{b}_r = \mathbf{b}m_2, \mathbf{c}_r = \mathbf{c}m_3\}$. The total number of symmetry operators for the replicated super cell is given by $n_r = n_s \times n_u$. The n_r fractional symmetry operators can be generated from the n_s fractional symmetry operators of the space group as

$$(\mathbf{R}_{ijkl}, \mathbf{t}_{ijkl}) = \left(\mathbf{R}_i, \left\{\frac{t_{i,1}+j}{m_1}, \frac{t_{i,2}+k}{m_2}, \frac{t_{i,3}+l}{m_3}\right\}\right) \quad (2)$$

and the Cartesian symmetry operators generated by

$$(\boldsymbol{R}_{ijkl}, \boldsymbol{t}_{ijkl}) = (\boldsymbol{R}_i, \boldsymbol{t}_i + \mathbf{n}_{jkl}) \quad (3)$$

where $i = 1, ..., n_s$, $j = 0, ..., m_1 - 1$, $k = 0, ..., m_2 - 1$ and $l = 0, ..., m_3 - 1$. In both the fractional and Cartesian cases each replicates super cell rotation matrix is equal to a rotation matrix of the space group. The fractional translation vector of each symmetry operator is scaled down in proportion to the number of replicated unit cells in each dimension such that enumeration of the n_r unit cells over the indices j, k, and l fills the replicates super cell. Similarly, the lattice vector $\mathbf{n}_{jkl} = j\mathbf{a} + k\mathbf{b} + l\mathbf{c}$ is added to the original Cartesian translation vectors.

C. Electrostatics under Periodic Boundary Conditions. There are two distinct physical pictures associated with lattice summation that have subtle but important differences.[14] The Ewald picture is based on an infinite lattice, which can be defined mathematically even though it is physically unrealizable. In this case the electrostatic potential obeys periodic boundary conditions, specifically $\Phi(\mathbf{r}) = \Phi(\mathbf{r} + n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c})$ for any set of integers $\{n_1,n_2,n_3\}$. The second physical picture is based on embedding a finite spherical lattice of unit cells inside a continuum dielectric and then taking the limit as its radius is increased to infinity. In this case, the electrostatic potential does not obey periodic boundary conditions due to two additional fields. The first is proportional to the dipole moment of the unit cell Φ_{dipole}, and the second is due to the reaction field Φ_{RF} of the dielectric medium that is induced by the spherical lattice.[43−46] If the dielectric of the surrounding medium is a vacuum, then there can be no reaction field, but the cell dipole field remains. On the other hand, if the dielectric of the medium is infinite, then the continuum reaction field cancels the dipole field. However, the physical picture of an embedded spherical lattice, even under so-called tinfoil boundary conditions with an infinite dielectric, is not equivalent to an infinite lattice and true periodic boundary conditions.[12]

We note that sampling from an embedded spherical lattice is conceptually problematic. Consider equivalent electrostatic charges separated by a lattice vector, for example, at locations $\mathbf{r}_i$ and $\mathbf{r}_i + n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$, that experience different dipole and/or reaction field forces. During a simulation only the coordinates of the central cell are explicitly propagated. In effect, the central unit cell and a unit cell on the edge of the embedded sphere are constrained to sample equivalent ensembles. Although both boundary conditions have limitations, in this work we focus on the Ewald infinite lattice picture and will not include further discussion of the embedded spherical lattice.

D. Asymmetric Unit Lattice Summation. To motivate the notation consider the electric potential at atom j located at $\mathbf{r}_j$ due to a collection of n_c point charges around atom i located at $\mathbf{r}_i$, each with a magnitude and position denoted by c_k and $\mathbf{r}_k$:

$$V(\mathbf{r}_j) = \frac{1}{4\pi\varepsilon_0}\sum_{k=1}^{n_c}\frac{c_k}{|\mathbf{r}_{ij} - \mathbf{r}_k|} \quad (4)$$

where $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$, and the Coulomb constant $1/4\pi\varepsilon_0$ will be neglected for convenience throughout the rest of the article. The potential can be expanded in a Taylor series to give

$$V(\mathbf{r}_j) = \sum_{k=1}^{n_c} c_k\left(1 + r_{k,\alpha}\nabla_{i,\alpha} + \frac{1}{2}r_{k,\alpha}r_{k,\beta}\nabla_{i,\alpha}\nabla_{i,\beta}\right)\frac{1}{r_{ij}} \quad (5)$$

where $\nabla_{i,\alpha}$ is one component of the del operator acting at $\mathbf{r}_i$, $\alpha \in \{x,y,z\}$ and the Greek subscripts $\{\alpha,\beta,\gamma,\delta,...\}$ represent use of the Einstein summation convention for summing over tensor elements.[47] The monopole, dipole, and traceless quadrupole moments are defined as

$$\begin{aligned} q_i &= \sum_{k=1}^{n_c} c_k, \\ d_{i,\alpha} &= \sum_{k=1}^{n_c} c_k r_{k,\alpha}, \\ \Theta_{i,\alpha\beta} &= \sum_{k=1}^{n_c} c_k\left(\frac{3}{2}r_{k,\alpha}r_{k,\beta} - \frac{1}{2}r_k^2\delta_{\alpha\beta}\right) \end{aligned} \quad (6)$$

where use of a traceless quadrupole is permitted since the potential satisfies the Laplace equation. Based on eq 6, we substitute the multipole moments back into the potential of eq 5 and define the multipolar operator L_i:

$$\begin{aligned} V(\mathbf{r}_j) &= L_i\left(\frac{1}{r_{ij}}\right), \\ L_i &= q_i + \mathrm{d}_{i,\alpha}\nabla_{i,\alpha} + \frac{1}{3}\Theta_{i,\alpha\beta}\nabla_{i,\alpha}\nabla_{i,\beta} \end{aligned} \quad (7)$$

Similarly, we can define the potential at $\mathbf{r}_i$ due to the multipole at $\mathbf{r}_j$ using multipolar operator L_j:

$$\begin{aligned} V(\mathbf{r}_i) &= L_j\left(\frac{1}{r_{ij}}\right), \\ L_j &= q_j - \mathrm{d}_{j,\alpha}\nabla_{i,\alpha} + \frac{1}{3}\Theta_{j,\alpha\beta}\nabla_{i,\alpha}\nabla_{i,\beta} \end{aligned} \quad (8)$$

where the sign difference between the multipolar operators is due to the relationship $\nabla_i = -\nabla_j$ for the function $|\mathbf{r}_i - \mathbf{r}_j|$. In the case of the AMOEBA force field, multipole coefficients are derived from electronic structure calculations on model chemical compounds using distributed multipole analysis (DMA).[23,48,49]

The potential energy U of the n_a permanent multipoles that make up the asymmetric unit is given by the lattice summation:

$$U = \frac{1}{2}\frac{1}{n_s}\sum_{n}^{*}\sum_{s_i=1}^{n_s}\sum_{s_j=1}^{n_s}\sum_{i=1}^{n_a}\sum_{j=1}^{n_a} L_i(\mathbf{R}_{s_i})L_j(\mathbf{R}_{s_j})\frac{1}{|\mathbf{x}|} \quad (9)$$

where $\mathbf{x} = \boldsymbol{R}_{s_i}\mathbf{r}_i + \boldsymbol{t}_{s_i} - (\boldsymbol{R}_{s_j}\mathbf{r}_j + \boldsymbol{t}_{s_j}) + \mathbf{n}$, the outer sum is over all lattice vectors $\mathbf{n} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$, the second and third sums are over the n_s symmetry operators of the space group that operate on sites i and j, respectively, and the inner sums are over the n_a multipole sites of the asymmetric unit. The asterisk denotes

skipping (or scaling) masked interaction pairs $(i, j) \in M$ in the list M and omission of self-interactions defined by $i = j$ for the central unit cell ($\mathbf{n} = 0$) and the identity symmetry operators ($s_i = s_j = 1$). A common example of masking is to omit the interaction between atoms that are covalently bonded. The multipolar operators L_i and L_j now include a Cartesian rotation matrix from a symmetry operator that rotates the multipole moments into the symmetry mate orientation:

$$L_i(\mathbf{R}) = q_i + (\mathbf{R}\mathbf{d}_i)_\alpha \nabla_{i,\alpha} + (\mathbf{R}\Theta_i\mathbf{R}^t)_{\alpha\beta}\nabla_{i,\alpha}\nabla_{i,\beta}\frac{1}{3} \quad (10)$$

and

$$L_j(\mathbf{R}) = q_j - (\mathbf{R}\mathbf{d}_j)_\alpha \nabla_{i,\alpha} + (\mathbf{R}\Theta_j\mathbf{R}^t)_{\alpha\beta}\nabla_{i,\alpha}\nabla_{i,\beta}\frac{1}{3} \quad (11)$$

Finally, division by two in eq 9 avoids double counting each interaction, and division by n_s converts from the unit cell energy to the asymmetric unit energy.

E. Asymmetric Unit Ewald Summation. Ewald summation[12] is based on multiplication of each term in eq 9 by a convergence function $\mathrm{erfc}(\beta|\mathbf{x}|)$ and then by $1 - \mathrm{erfc}(\beta|\mathbf{x}|) = \mathrm{erf}(\beta|\mathbf{x}|)$ to give

$$\begin{aligned} U &= U_{\mathrm{real}} + U_{\mathrm{recip}} \\ &= \frac{1}{2}\frac{1}{\mathrm{n}_s}{\sum_n}^{*}\sum_{s_i=1}^{n_s}\sum_{s_j=1}^{n_s}\sum_{i=1}^{n_a}\sum_{j=1}^{n_a} L_i(\mathbf{R}_{s_i})L_j(\mathbf{R}_{s_j})\frac{\mathrm{erfc}(\beta|\mathbf{x}|)}{|\mathbf{x}|} \\ &\quad + \frac{1}{2}\frac{1}{n_s}{\sum_n}^{*}\sum_{s_i=1}^{n_s}\sum_{s_j=1}^{n_s}\sum_{i=1}^{n_a}\sum_{j=1}^{n_a} L_i(\mathbf{R}_{s_i})L_j(\mathbf{R}_{s_j})\frac{\mathrm{erf}(\beta|\mathbf{x}|)}{|\mathbf{x}|} \end{aligned} \quad (12)$$

where β is the Ewald convergence parameter. The first summation U_{real} is rapidly decreasing and may be evaluated in real space by ignoring all terms outside of a cutoff radius r_c, which is typically chosen between 7 and 9 Å. An appropriate β can be determined by satisfying $\mathrm{erfc}(\beta r_c)/r_c < \varepsilon_{\mathrm{real}}$ at the cutoff for a target error tolerance $\varepsilon_{\mathrm{real}}$. The second term is smooth, periodic, and rapidly decreasing in reciprocal space if masked and if self-interactions are added back, which is discussed further below. The physical picture is that a 3D Gaussian charge density has been added and then subtracted at the location of each point charge (or appropriate gradients of the Gaussian density for dipole, quadrupole, or higher order moments). As the Ewald convergence parameter β is increased, for example, to satisfy the target error tolerance for a small real space cutoff, relatively higher frequencies must be included in the reciprocal space sum. In this manner β can be used to tune the relative rate of convergence of the two sums.

F. Real Space Summation. The real space sum in eq 12 can be simplified to

$$U_{\mathrm{real}} = \frac{1}{2}\sum_{s_j=1}^{n_r}{}^{*}\sum_{i=1}^{n_a}\sum_{j=1}^{n_a} L_i(\mathbf{I})L_j(\mathbf{R}_{s_j})\frac{\mathrm{erfc}(\beta|\mathbf{r}_i - (\mathbf{R}_{s_j}\mathbf{r}_j + \boldsymbol{t}_{sj})|)}{|\mathbf{r}_i - (\mathbf{R}_{s_j}\mathbf{r}_j + \boldsymbol{t}_{sj})|} \quad (13)$$

where the asterisk now indicates that $i = j$ interactions are neglected and masked interactions $(i,j) \in M$ are respected for the identity symmetry operator $s_j = 1$. A replicates super cell and n_r symmetry operators are required for a unit cell whose smallest width is less than half the real space cutoff.[42] In this way all interactions within the real space cutoff are treated consistently via application of the minimum image convention using the replicates super cell basis vectors $\{\mathbf{a}_r,\mathbf{b}_r,\mathbf{c}_r\}$. The sum over lattice vectors $\mathbf{n}$ can be removed, since any lattice vector with length greater than zero produces interactions outside of the real space cutoff distance. When a replicates super cell is not required, then n_r is equal to n_s unit cell space group symmetry operators, and application of the minimum image convention is based on the unit cell basis vectors $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. Since the real space energy for each copy of the asymmetric unit is equal, the sum over symmetry operators for multipole site i may be removed, and division by the number of space group operators is unnecessary.

We emphasize an important difference between eq 13 and the analogous eq (2.13) of Sagui et al. due to the inclusion of symmetry operators.[16] The summations over i and j are reduced compared to a P1 unit cell from n_{P1} multipole sites to $n_a = n_{\mathrm{P1}}/n_s$. If the asymmetric unit is large relative to the real space cutoff, then the reduction in terms relative to a calculation in P1 approaches a factor of n_s. When the real space work is numerically expensive relative to the reciprocal space work, as for multipolar force fields or electronic structure calculations, the speedup for the overall calculation approaches n_s.

Details for applying the multipolar operator in the case of the AMOEBA potential can be found in the Appendix of the work by Ren and Ponder and will not be repeated here.[5] Briefly, the derivation of real space Ewald summation for multipoles presented by Smith[50] is modified by a Thole damping function[51] at short range when computing polarization interactions for AMOEBA. However, we recommend consideration of the McMurchie−Davidson recursion,[52,53] as presented by Sagui et al.[16] when moments above quadrupole order are considered.

G. Reciprocal Space Summation. The reciprocal space summation requires adding and then subtracting the self-energy U_{self} and masked interaction energy U_{mask} that were excluded from eq 12 to give

$$U_{\mathrm{recip}} = U_{\mathrm{Periodic}} - U_{\mathrm{self}} - U_{\mathrm{mask}} \quad (14)$$

The term U_{periodic} is smooth, periodic, and its Fourier transform is given by[13]

$$\hat{U}_{\mathrm{periodic}} = \frac{1}{n_s}\frac{1}{2\pi V}\sum_{h\neq 0}\frac{\exp(-\pi^2 s^2/\beta^2)}{\mathbf{s}^2}\hat{F}(\mathbf{h})\hat{F}(-\mathbf{h}) \quad (15)$$

where $\mathbf{s} = \mathbf{h}^t\mathbf{A}^{-1}$ is the scattering vector, $\mathbf{h}$ contains the Miller indices of a Brag reflection, and $V = \mathbf{a}\cdot\mathbf{b} \times \mathbf{c}$ is the volume of the unit cell. The total multipolar structure factor,[16] including a summation over unit cell symmetry operators, is given by

$$\hat{F}(\mathbf{h}) = \sum_{i=1}^{n_a}\sum_{s=1}^{n_s}\hat{L}_j(\mathbf{s},\mathbf{R}_s)\exp[2\pi i\mathbf{h}\cdot(\mathbf{R}_s\mathbf{A}^t\mathbf{r}_j + \mathbf{t}_s)] \quad (16)$$

where the Fourier transform of the multipolar operator L_j is given by

$$\hat{L}_j(\mathbf{s},\mathbf{R}) = q_j + 2\pi i(\mathbf{R}\mathbf{d}_j^t)_\alpha s_\alpha - 4\pi^2(\mathbf{R}\Theta_j\mathbf{R}^t)_{\alpha\beta}s_\alpha s_\beta \quad (17)$$

Alternatively, symmetry operators can be applied in reciprocal space[54] to the structure factor for the asymmetric unit:

$$\hat{F}^A(\mathbf{h}) = \sum_{j=1}^{n_a}\hat{L}_j(\mathbf{s},\mathbf{I})\exp[2\pi i\mathbf{h}\cdot(\mathbf{r}_j^t\mathbf{A})] \quad (18)$$

based on the expression:

$$\hat{F}(\mathbf{h}) = \sum_{s=1}^{n_s} \hat{F}^A(\mathbf{R}_s^t\mathbf{h}) \exp(2\pi i \mathbf{h}\cdot\mathbf{t}_s) \tag{19}$$

where the term $\exp(2\pi i\mathbf{h}\cdot\mathbf{t}_s)$ is due to the translational part of the symmetry operator.

H. Ewald Self- and Masked Interactions. The self-interaction terms can be determined by taking the limit of $f(r) = \mathrm{erf}(\beta r)/r$ and its partial derivatives, as specified by the multipolar operators in eq 15 in the limit $r \rightarrow 0$ to give

$$\begin{aligned} \lim_{r\rightarrow 0} f(r) &= \frac{2\beta}{\sqrt{\pi}} \\ \lim_{r\rightarrow 0}[\nabla_\alpha(-\nabla_\alpha)f(r)] &= \frac{4\beta^3}{3\sqrt{\pi}} \\ \lim_{r\rightarrow 0}[\nabla_\alpha^2\nabla_\beta^2(1/3)^2 f(r)] &= \frac{8\beta^5}{45\sqrt{\pi}} \\ \lim_{r\rightarrow 0}[\nabla_\alpha^4(1/3)^2 f(r)] &= \frac{24\beta^5}{45\sqrt{\pi}} \end{aligned} \tag{20}$$

Based on the results of eq 20 the total self-interaction energy U_{self} that must be removed from U_{recip} is given by

$$U_{\mathrm{self}} = \frac{1}{2}\sum_{i=1}^{n_a}\frac{2\beta}{\sqrt{\pi}}q_i^2 + \frac{4\beta^3}{3\sqrt{\pi}}d_{i,\alpha}^2 + \frac{16\beta^5}{45\sqrt{\pi}}\Theta_{i,\alpha\beta}^2 \tag{21}$$

which is consistent with the result of Aguado et al.[55] The quadrupole self-interaction term is based on intermediate steps that depend on it being traceless and symmetric. Since they have not been presented previously, we provide these steps below. From eq 20 the self-interaction for an element of the quadrupole trace is due to the interaction with itself and the other two trace elements:

$$\begin{aligned} &\frac{\beta^5}{45\sqrt{\pi}}[24\Theta_{\alpha\alpha}^2 + 8\Theta_{\alpha\alpha}(\Theta_{\beta\beta}+\Theta_{\gamma\gamma})] \\ &= \frac{\beta^5}{45\sqrt{\pi}}[24\Theta_{\alpha\alpha}^2 + 8\Theta_{\alpha\alpha}(-\Theta_{\alpha\alpha})] = \frac{16\beta^5}{45\sqrt{\pi}}\Theta_{\alpha\alpha}^2 \end{aligned} \tag{22}$$

while the self-interaction for an off-diagonal element is due to the interaction with itself and the symmetric element:

$$\frac{8\beta^5}{45\sqrt{\pi}}\Theta_{\alpha\beta}(\Theta_{\alpha\beta}+\Theta_{\beta\alpha}) = \frac{16\beta^5}{45\sqrt{\pi}}\Theta_{\alpha\beta}^2 \tag{23}$$

Masked terms $(i,j) \in M$ were easily accounted for in the real space sum but included at full strength in the Fourier sum to enforce exact periodicity. The overcounting can be removed by subtracting the real space sum over masked interactions within the asymmetric unit given by

$$U_{\mathrm{mask}} = \frac{1}{2}\sum_{i,j\in \mathrm{M}} L_i(\mathbf{I})L_j(\mathbf{I})\frac{\mathrm{erf}(\beta|\mathbf{r}_i-\mathbf{r}_j|)}{|\mathbf{r}_i-\mathbf{r}_j|} \tag{24}$$

I. PME Reciprocal Space Summation. Instead of direct summation of structure factors, they can be computed via B-spline interpolation onto a discrete grid followed by 3D FFT. This is analogous to the method used to compute crystallographic structure factors, with the notable differences that point multipoles are interpolated at grid points using B-splines that have finite support, whereas Gaussian form factors are explicity evaluated at grid points and have infinite support necessitating truncation outside of a cutoff.[56] Smooth PME interpolates multipoles to a finite set of nearby grid points using cardinal B-splines $\theta_p(u)$ of order p as described originally by Essmann et al. for fixed charge models[13] and later extended to higher order moments.[15,16] The first order cardinal B-spline $\theta_1(u)$ is defined as the characteristic function of $[0,1]$ and higher orders recursively as the convolution product:

$$\theta_k = \theta_{k-1} * \theta_1 \tag{25}$$

The support of $\theta_k(u)$ is compact and given by $[0,k]$. The error of the PME approximation can be systematically reduced via higher order B-splines in tandem with finer grids.

The complex exponential in eq 16 may be expanded to

$$\exp(2\pi i\mathbf{h}\cdot\mathbf{u}_i) = \exp(2\pi i h u_{i,1})\exp(2\pi i k u_{i,2})\exp(2\pi i l u_{i,3}) \tag{26}$$

where $\mathbf{u}_i$ are the fractional coordinates of site i after application of the symmetry operator s_i as given by $\mathbf{u}_i = \mathbf{R}_{s_i}\mathbf{A}^t\mathbf{r}_i + \mathbf{t}_{s_i}$. The Euler exponential spline s_b is then used to interpolate each complex exponential[13] of eq 26 as

$$\begin{aligned} \exp(2\pi i h_\alpha u_{i,\alpha}) &\approx s_b(h_\alpha, u_{i,\alpha}) \\ &= b_\alpha\left(\frac{h_\alpha}{N_\alpha}\right)\sum_{k=-\infty}^{\infty}\theta_p(N_\alpha u_{i,\alpha}-k) \\ &\quad\cdot\exp\left(2\pi i\frac{h_\alpha}{N_\alpha}k\right) \end{aligned} \tag{27}$$

for grid dimension N_α and coefficients $b_\alpha(h_\alpha/N_\alpha)$ given by

$$b_\alpha\left(\frac{h_\alpha}{N_\alpha}\right) = \frac{\exp[2\pi i(p-1)h_\alpha/N_\alpha]}{\sum_{k=0}^{p-2}\theta_p(k+1)\cdot\exp(2\pi i k h_\alpha/N_\alpha)} \tag{28}$$

The fractional grid array that includes the contributions of all multipoles is given by

$$Q(\mathbf{k}) = \sum_{s_i=1}^{n_s}\sum_{i=1}^{n_a}\sum_n \hat{L}_i(\mathbf{R}_{si})\begin{bmatrix} & \theta_p(u_{i,1}N_1 - k_1 - n_1N_1) \\ \times & \theta_p(u_{i,2}N_2 - k_2 - n_2N_2) \\ \times & \theta_p(u_{i,3}N_3 - k_3 - n_3N_3) \end{bmatrix} \tag{29}$$

where $\mathbf{k} = \{k_1,k_2,k_3\}$ is a point of the $\mathbf{N} = \{N_1,N_2,N_3\}$ sized 3D grid and $\mathbf{n} = \{n_1,n_2,n_3\}$ indicates a sum over all integer triples. The inner sum is actually finite due to local support of each B-spline. The discrete Fourier transform of eq 29 is given by

$$\begin{aligned} \hat{Q}(\mathbf{h}) = &\sum_{k_1=0}^{N_1-1}\sum_{k_2=0}^{N_2-1}\sum_{k_3=0}^{N_3-1} Q(\mathbf{k}) \\ &\exp[2\pi i\mathbf{h}\cdot(k_1/N_1, k_2/N_2, k_3/N_3)] \end{aligned} \tag{30}$$

and the analytic structure factor of eq 16 can be approximated as

$$\hat{F}(\mathbf{h}) = \begin{cases} B(\mathbf{h})\hat{Q}(\mathbf{h}) & -\frac{1}{2} \le \left\{\frac{h}{N_1},\frac{k}{N_2},\frac{l}{N_3}\right\} < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

where

$$B(\mathbf{h}) = b_1\left(\frac{h}{N_1}\right)\cdot b_2\left(\frac{k}{N_2}\right)\cdot b_3\left(\frac{l}{N_2}\right) \tag{32}$$

The periodic portion of the reciprocal energy is then computed as before using eq 15. In our implementation the principle quantity of interest is the reciprocal electrostatic potential, and its gradients at each multipole site within the asymmetric unit, given by[16]

$$\begin{aligned}\varphi_{\mathrm{rec}}(\mathbf{r}_i) &= \frac{1}{\pi V}\sum_{h\neq 0}\frac{\exp(-\pi^2\mathbf{h}^2/\beta^2)}{\mathbf{h}^2}s_b(-h,u_{i,1})s_b(-k,u_{i,2})s_b(-l,u_{i,3})\\ &\cdot\hat{F}(\mathbf{h}) = \sum_{\mathbf{n}}\theta_p(N_1u_{i,1}-n_1)\theta_p(N_2u_{i,2}-n_2)\theta_p(N_3u_{i,3}-n_3)\\ &\cdot(G{*}Q)(\mathbf{n})\end{aligned} \tag{33}$$

where the second equality follows from Parseval's identity with Q given by eq 29 and G defined as the inverse discrete Fourier transform of a generalized influence function:

$$\hat{G}(\mathbf{h}) = \begin{cases}\frac{1}{\pi V}|B(\mathbf{h})|^2\frac{\exp(-\pi^2s^2/\beta^2)}{s^2} & -\frac{1}{2}\le\left\{\frac{h}{N_1},\frac{k}{N_2},\frac{l}{N_3}\right\}<\frac{1}{2},\mathbf{s}\neq 0\\ 0 & \text{otherwise}\end{cases} \tag{34}$$

The convolution of the pair potential G and multipole array Q gives the potential on the grid at $\mathbf{n}$, which is evaluated at a finite number of nonzero grid points in real space due to the finite support of the B-splines. The potential only needs to be evaluated for atoms within the asymmetric unit, which speeds up this part of the calculation by a factor of the number of space group symmetry operators. Gradients of the potential are found by taking gradients of θ_p as described in earlier work.[16]

In reciprocal space, the convolution $C(\mathbf{n}) = (G{*}Q)(\mathbf{n})$ becomes a simple multiplication for each structure factor:

$$\hat{C}(\mathbf{h}) = \begin{cases}\hat{G}(\mathbf{h})\hat{Q}(\mathbf{h}) & -\frac{1}{2}\le\left\{\frac{h}{N_1},\frac{k}{N_2},\frac{l}{N_3}\right\}<\frac{1}{2}\\ 0 & \text{otherwise}\end{cases} \tag{35}$$

Performing the inverse discrete 3D FFT on $\hat{C}$ generates the desired convolution product

$$C(\mathbf{n}) = \frac{1}{N_1N_2N_3}\sum_{k_1=0}^{N_1-1}\sum_{k_2=0}^{N_2-1}\sum_{k_3=0}^{N_3-1}\hat{C}(\mathbf{k})\exp[-2\pi i\mathbf{n}\cdot(k_1/N_1,k_2/N_2,k_3/N_3)] \tag{36}$$

For optimal computational performance, it is advantageous to view the reciprocal space portion of the calculation in terms of a *single overall convolution operation*, rather than three serial steps as described above:

(1) 3D FFT given in eq 30.
(2) Reciprocal space multiplication given in eq 35.
(3) 3D inverse FFT in eq 36. This idea will be emphasized in the following section on our parallel implementation.

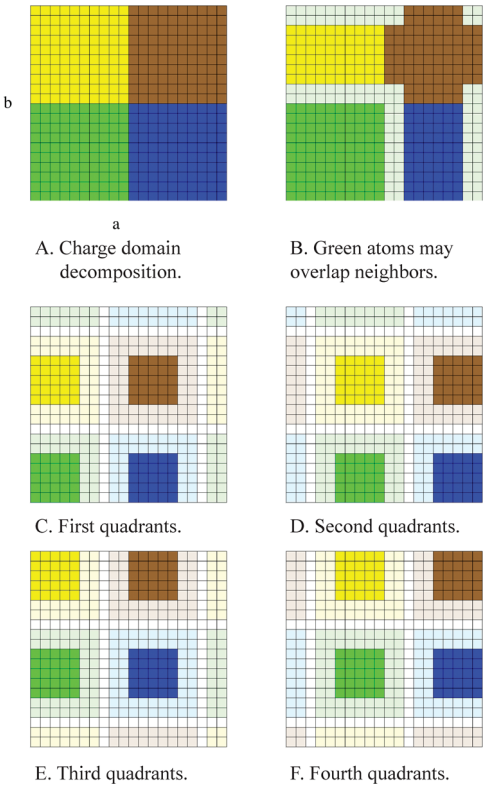

Figure 1. Panel A shows one face of a domain decomposition for a 20 × 20 × 20 grid of source density in fractional coordinates. All atoms within a color-coded domain are assigned to the same compute core. Panel B highlights in light green the grid region that is not assigned to the green core but receives source density from atoms in the green region (the effect of PBC is included). This is based on quintic b-Splines whose support region is a 5 × 5 × 5 grid, although the method is easily adjusted to other support requirements. The yellow, brown, and blue cores must not attempt to modify the source density values at light-green grid points, while the green core is modifying them. Panels C−F demonstrate further subdivision of each CPU region into quadrants. In each of four synchronized steps, atoms within the active green, yellow, brown, and blue quadrant have their source density spread to the grid. Grid regions colored light green, light yellow, light brown, and light blue represent the maximum extent of source density spreading by their dark central quadrant. The white outline separating the maximum extent of support indicates that no two cores will require the same grid point during a step of the procedure. Slow atomic operations in software and hardware specific APIs are replaced by a few high-level thread synchronizations.

III. PARALLEL IMPLEMENTATION

We now focus on the shared memory parallelization of the reciprocal space portion of PME as implemented in FFX. This portion of the calculation is the limiting factor for force field energy evaluations for large biomolecular crystals due to $N\cdot\log(N)$ scaling of the FFT. The real space portion of our algorithm has also been parallelized, however, our view is that the combination of N-body summations with symmetry operators merits a separate, self-contained treatment. The reason is most zonal schemes or spatial decompositions assume nearly uniform particle density over the unit cell, which is not the case after removing redundant copies of the asymmetric unit.[57]

First we discuss a general domain decomposition scheme for spreading source density onto the 3D FFT grid, as described by eq 29. Then we present two parallelization strategies for the

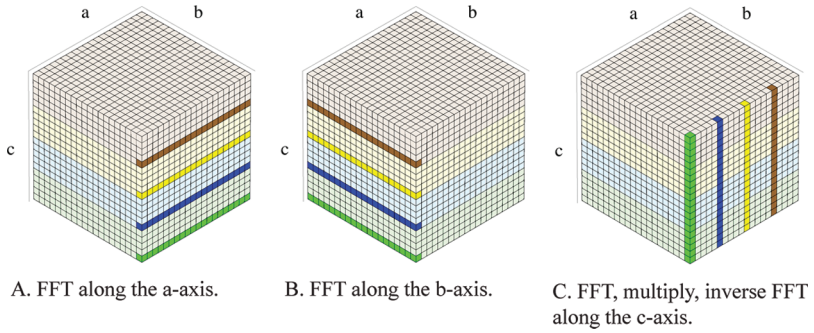

Figure 2. Panels A−C depict partitioning of the 20 × 20 × 20 source grid into *ab*-planes that are distributed among four cores. Panels A and B represent 1D FFTs along the *a*-axis and *b*-axis, respectively. For these operations the required data is stored in memory assigned to the core doing the transform. Panel C shows the final 1D FFT that requires grid values that are distributed in memory across all cores. Before a thread moves on to the next row, the reciprocal space multiplications and inverse 1D FFT are performed to minimize cache misses or cache interference. This optimization is not possible for algorithms based on completing the 3D FFT for the entire grid before doing the reciprocal space multiplication. The final *a*-axis and *b*-axis inverse 1D FFTs are not shown.

Table 1. Shown Are Timings for Our 3D FFT Routines Based on Either Real or Complex Input Data for Transforms Sizes of 64^3, 128^3, and 256^3 [a]

	64^3				128^3				256^3			
method	1	8	X	CUDA	1	8	X	CUDA	1	8	X	CUDA
R	0.034	0.009	3.7		0.238	0.056	4.2		2.291	0.321	7.1	
R Conv.	0.028	0.006	4.4		0.228	0.040	5.6		2.140	0.248	8.6	
C	0.061	0.014	4.3	*0.003*	0.545	0.084	6.4	*0.017*	7.715	1.004	7.7	*0.224*
C Conv.	0.055	0.010	5.5		0.492	0.067	7.4		6.886	0.648	10.6	

[a] The real (R) and complex (C) timings are for the sum of 3D FFT and inverse 3D FFT called sequentially. For the real convolution (R Conv.) and complex convolution (C Conv.), the timing includes the 3D FFT, reciprocal space multiplication, and inverse 3D FFT treated a *single operation*. For each combination, both one and eight cores were tested, and the speed-up is shown in the column labeled X. The timings for CUDA are *italicized* because they were done in single precision, while all others calculations are double precision. All timings are in seconds. The CUDA library does not include *single operation* convolutions (R Conv. and C Conv.) and our work did not require implementation of the real CUDA sequential approach (R) so these table entries are blank. Speedups greater than 8 result from cache effects.

reciprocal space convolution. The key to the first algorithm is viewing the convolution as a single operation and has been parallelized for a shared memory JRE. The second algorithm is currently based on the nVidia CUDA language and its 3D FFT library, which necessitates viewing the convolution as a sequential series of three operations, as described above.

A. Spreading Source Density onto the 3D Grid. A key issue in parallelization of charge density spreading is that two threads of execution cannot be allowed to modify the value of any grid point concurrently. For example, consider the simplest possible spatial decomposition, namely two domains of equal size separated by two parallel planes (one plane is a periodic boundary and the other is parallel to it). Atoms that are near a plane will spread source density into both subdomains. Now consider dividing both subdomains in half by two additional planes, parallel to the first two, to give four total subdomains of equal size. As long as each subdomain dimension is as large as the support of the atomic source density, for example, five grid points in each dimension for quintic b-Splines, then it is guaranteed that subdomains that do not share a plane do not interpolate multipoles into each other. In our trivial example, therefore, threads may operate simultaneously on regions one and three without requiring access to the same grid point. When the two threads of execution complete regions one and three, they synchronously continue to regions two and four. More generally, there may be $d_\alpha = N_\alpha/b_\alpha$ subdomains and $p_\alpha = d_\alpha/2$ subdomain pairs along the

Table 2. Presented Are the Timings and Speed-up for the Evaluation of the Acetamide Crystal Structure Using the AMOEBA Force Field and PME Electrostatics[a]

simulation cell	acetamide molecules	time (sec)	speed-up
2 × 2 × 2 P1	144	0.584	1.0
P1	18	0.114	5.1
H3c	1	0.016	36.5

[a] The 2 × 2 × 2 replicated unit cell avoids the need for an explicit replicates algorithm, since the super cell edges are greater than twice the real space cutoff. The combination of space group and replicates operators in FFX gives a speed-up for the acetamide crystal of more than 36× relative to the 2 × 2 × 2 replicated unit cell without parallelization.

α-axis $\alpha \in \{a,b,c\}$ with grid dimension N_α and support requirement b_α. In two dimensions, there may be at most $d_{\alpha,\beta} = (N_\alpha/b_\alpha)(N_\beta/b_\beta)$ subdomains and $q_{\alpha,\beta} = d_{\alpha,\beta}/4$ subdomain quartets requiring 3 synchronization steps to avoid sharing planes, as shown in Figure 1. Finally, in three dimensions, there may be at most $d_{a,b,c} = (N_a/b_a)(N_b/b_b)(N_c/b_c)$ subdomains and $o_{a,b,c} = d_{a,b,c}/8$ octets requiring 7 synchronization steps to avoid sharing planes. Note that each division above must be done separately to ensure an even result along each axis.

B. Reciprocal Space Convolution. Parallelization of the reciprocal space convolution is of critical importance to the

parallel scaling of PME electrostatics. We briefly discuss our CPU parallelization scheme and its relative merits and also refer to more comprehensive and focused presentations.[58] The 3D transform is decomposed into 1D transforms along each axis, as shown in Figure 2. First N_b times N_c transforms of length N_a are performed along the a-axis. Then, $N_a \times N_c$ transforms of length N_b are performed along the b-axis. Finally, $N_a \times N_b$ transforms of length N_c are performed along the c-axis. If the data is packed in a 1D array in memory, with dimension a varying most quickly, dimension b varying second most quickly, and dimension c varying most slowly, then the transforms along the c-axis require the most severely nonlocal memory accesses. With this in mind, ab-planes are divided equally among available CPUs, and the first two sets of 1D transforms are very memory efficient. Transforms along the c-axis are then divided equally among available CPUs. Before each transform a thread-local array of length N_c is packed contiguously with values otherwise separated by $N_a \times N_b$ complex values in memory. For PME, the point-wise reciprocal multiply of eq 35 should be performed and the inverse FFT along the c-axis done immediately. Finally, the local result is copied back into the global array, before the thread moves on to its next c-axis transform.

Optionally, the reciprocal space calculation can be accelerated using the CUDA API, which does not include a convolution operation. Instead, the forward 3D FFT, reciprocal multiplication, and inverse 3D FFT are done sequentially. The CPUs still perform the real space calculation while the PME grid is transferred to the GPU, the convolution is performed, and finally the result is transferred back to main memory. The transfer time becomes insignificant for large 3D grids so that our overall algorithm scales $N \cdot \log(N)$ on the GPU (Table 1). Since we are currently using single precision for optional GPU acceleration, a discussion of single vs double precision is given in Section D of the Supporting Information.

IV. APPLICATIONS

There are a significant variety of applications where explicit inclusion of symmetry operators within PME described in this work and implemented within FFX may play an essential role. At the small organic molecule end of the spectrum is ab initio crystal structure prediction and solubility estimation.[59,60] For example, the pharmaceutical industry is especially interested in polymorph prediction, where each polymorph can have different physical properties and bioavailability. Consider the case of acetaminophen, which crystallizes in three polymorphs.[61] At the other end of the spectrum is the refinement of large macromolecular structures at low resolution where de novo model building can be problematic without previously determined high-resolution substructures.[62,63] This work represents a first step toward demonstrating that the prior chemical information contained within the AMOEBA force field can be used to improve macromolecular models from refinement with medium- to low-resolution data sets.

A. Replicates for Small Molecule Crystals. Acetamide is an important model compound when developing a biomolecular force field and forms a crystal at standard temperature and pressure. For these reasons it was chosen to demonstrate the application of our space group plus replicates PME implementation of the AMOEBA force field to small organic crystals(Table 2). For comparison, a recent application of the fixed charge CHARMM force field to predict the crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl)

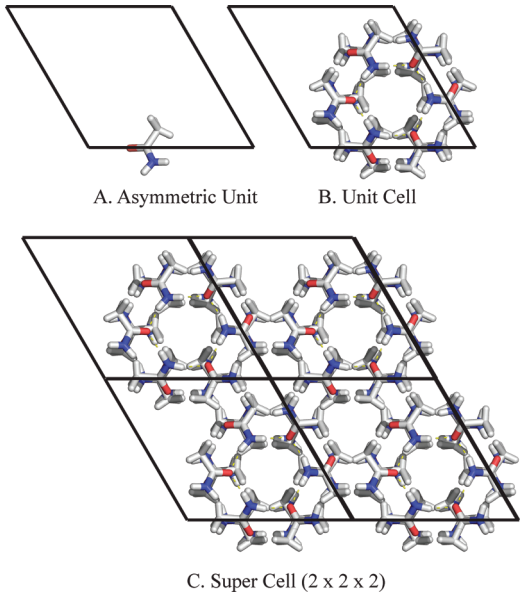

Figure 3. Shown in Panel A is the acetamide asymmetric unit (space group $H3c$). Panel B shows the unit cell after expansion to P1, which contains 18 molecules in identical chemical environments. Finally, Panel C shows a super cell ($2 \times 2 \times 2$) whose edge lengths are each at least twice the pair wise cutoff distance. Calculations of AMOEBA energies and gradients for the asymmetric unit are accelerated by a factor of more than 36× relative to a code without replicates operators and 7× relative to a code with replicates but without space group symmetry operators.

acetamide required expansion to P1 followed by the creation of at least two copies of the unit cell in each dimension.[64] Therefore, $2^3 \cdot N_{symm} - 1$ more copies of the asymmetric unit than necessary contribute to the computation of the asymmetric unit potential energy. In contrast, the algorithm presented here allows the calculation to be done on only the asymmetric unit. This is a useful step forward in terms of both force field quality (fixed charges replaced by polarizable atomic multipoles) and search efficiency. The small unit cell dimensions of the rhombohedral crystal (11.526, 11.526, 13.526, 90.0, 90.0, 120.0, space group $H3c$) are clearly not twice the distance of a converged real space cutoff for either PME or van der Waals energy. Since the unit cell contains 18 copies of acetamide, expansion to a $2 \times 2 \times 2$ P1 super cell contains 144 molecules, as shown in Figure 3.

B. Macromolecular X-ray Crystallography Refinement. Refinement of large macromolecular complexes, such as the ribosome, using a rigorous lattice summation method is a primary motivation for this work. For example, the importance of electrostatics to the mechanism of translation due to the interaction of divalent Mg^{2+} cations with the negatively charged phosphate backbone of rRNA, mRNA, and tRNA has been suggested by a recent 2.8 Å structure of *Thermus thermophilus*.[65] If the same structure could be solved to a higher resolution, perhaps below 2 Å, it is reasonable to expect further biological insights due to features that were unclear in the lower resolution electron density maps. As an example of such an improvement, consider the structure of the phenylalanine tRNA that was originally determined in the 1970s using ∼3 Å resolution data sets[66] but was recently improved via a data set at 1.93 Å.[67] The new, higher resolution model exhibits six additional metal sites, an average change in torsional angles of ∼40°, alternate sugar puckers, and extensive differences in water structure.

Table 3. Shown Are Timings for the Evaluation of the AMOEBA Potential Energy for 30 Crystallography Data Sets[a]

					time (seconds)				
PDB	space		number of atoms		UC	AU			
ID	group	b_{symm}	AU	UC	1 CPU	1 CPU	8 CPUs	8 CPUs + GPU	speed-up
1AV1	P212121	4	13 224	52 896	53.2	28.5	3.6	2.0	27.1
1A7B	P212121	4	5928	23 712	13.2	5.4	1.0	0.9	14.3
1BL8	C2	4	5898	23 592	33.0	18.2	3.6	2.4	13.6
1DP0	P212121	4	76 805	307 220	235.2	95.9	20.9	14.7	16.0
1ISR	P3221	6	7067	42 402	23.6	11.6	1.8	1.1	20.6
1JL4	P4322	8	8749	69 992	36.1	15.2	3.1	2.0	17.8
1JSE	P41212	8	88 347	706 776	971.2	195.7	35.1	20.6	47.1
1PGF	*I*222	8	17 770	142 160	117.0	59.2	10.5	5.2	22.6
1R5U	*I*222	8	56 797	454 376	345.9	192.4	24.7	10.0	34.8
1XDV	P212121	4	25 155	100 620	68.0	29.5	6.8	5.4	12.5
1XXI	P212121	4	55 778	223 112	110.8	57.3	8.2	5.3	20.9
1X8W	P41212	8	31 220	249 760	324.0	63.2	9.1	5.0	64.5
1YE1	P21212	4	9366	37 464	16.9	6.6	1.2	0.9	18.0
1YI5	C2221	8	20 961	167 688	94.3	36.9	7.9	5.9	15.9
1Z9J	P4222	8	12 807	102 456	114.3	63.5	10.9	4.6	25.1
2A62	P4122	8	4911	39 288	26.6	15.8	3.0	1.5	18.3
2BF1	P43212	8	4853	38 824	28.2	12.8	2.8	1.7	16.7
2FNP	P21	2	4201	8402	5.5	3.4	0.7	0.6	9.4
2I36	P3112	6	15 266	91 596	61.8	31.8	5.0	3.1	19.6
2J00	P212121	4	487 164	1 948 656	*3935.6*	1513.7	200.6	85.1	46.2
2QAG	P4322	8	8947	71 576	76.5	62.2	8.3	3.8	20.0
2QUK	*P*622	12	6235	74 820	72.0	29.6	4.5	3.0	24.1
2R4R	C2	4	10 068	40 272	26.3	13.7	2.3	1.6	16.1
2VKZ	P43212	8	171,819	1,374,552	*1036.8*	398.8	47.6	20.3	51.2
3BBW	P61	6	8865	53 190	36.1	18.2	3.6	1.8	19.7
3CRW	P212121	4	8305	33 220	19.2	7.0	1.2	1.0	18.7
3DMK	*I*222	8	32 758	262 064	153.0	69.3	9.5	4.9	31.2
3DU7	P65	6	27 491	164 946	141.3	82.2	14.4	8.3	17.0
3FFN	P4212	8	22 645	181 160	114.7	44.8	7.5	4.8	23.8
3HN8	P41212	8	13 395	107 160	77.5	37.1	5.7	4.0	19.4
mean		6.3	42 093	239 798	278.9	107.3	15.5	7.7	24.1

[a] The average number of atoms in the asymmetric unit (AU) is a 6.3 fold reduction compared to the average number of atoms in the unit cell (UC). The mean speed-up from space group symmetry, shared memory parallelization over 8 CPU cores, and a GPU coprocessor for the reciprocal space convolution is a factor of 24×.

We present timings for evaluation of the potential energy for 30 macromolecular crystals in Table 3. For example, the 3CRW asymmetric unit and unit cell are shown in Figure 4. The average number of atoms in the asymmetric unit (42 093) is already quite large relative to calculations that have been done with AMOEBA thus far.[8] The average number of atoms in the unit cell after expansion to P1 symmetry (239 798) is 6.3 fold higher still. For these timings and the optimizations described below, the van der Waals cutoff was set to 9.0 Å. A polynomial switch was used to smoothly turn off the van der Waals potential energy over a window width of 0.9 Å (starting at 8.1 Å). For PME, the real space cutoff was set to 7.0 Å, the Ewald convergence parameter set to 0.545, the B-spline order to 5 and a reciprocal space grid density of 1.2 grid points per Å, which are the currently recommended values for use with AMOEBA.[9] In some cases, the grid density was increased or decreased by no more than 10% to achieve power of 2 grid dimensions, which is currently maximally efficient for the CUDA FFT library. In the future, we anticipate OpenCL FFT libraries that suffer less performance degradation for nonpower of two sizes. The AMOEBA self-consistent field (SCF) was converged to a tolerance of 0.01 RMS Debye. This is also known as the *mutual* polarization model. For high-temperature simulated annealing,[68] the accuracy of *mutual* polarization may be unnecessary, and a *direct* polarization approximation can be used instead. Under the *direct* polarization approximation, the total field of the permanent multipoles influences the polarizable sites but not the field of the induced dipoles themselves (for details see Section A of the Supporting Information). For this reason, the *direct* approximation is about an order of magnitude faster than the true AMOEBA potential that requires SCF convergence.

By using space group symmetry, shared memory parallelization, and a GPU coprocessor for the reciprocal space convolution, the average time for an energy evaluation of these large

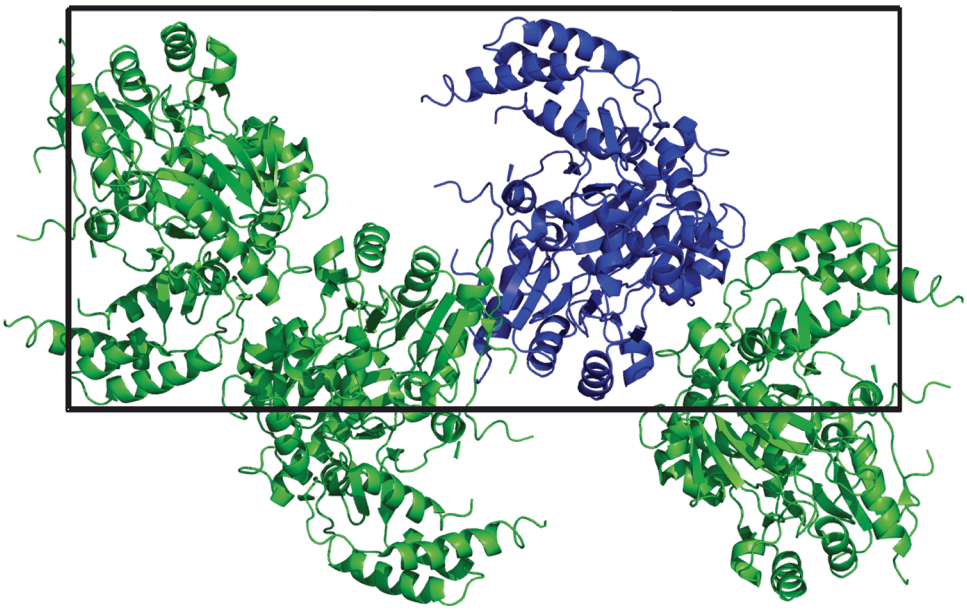

Figure 4. One of the smaller crystals (3CRW) is shown to given an impression of the number of real space interactions that are removed by considering only the asymmetric unit (blue) by eliminating redundant symmetry mates (green).

Table 4. Details for 10 Crystallography Data Used to Explore AMOEBA-Assisted Biomolecular X-ray Refinement Potential[a]

	res.	reported			FFX recalculated			geometry RMSD	
model	(Å)	R	R_{free}	$R_{free} - R$	R	R_{free}	$R_{free} - R$	bonds (Å)	angles (°)
1A7B	3.1	23.9	30.6	6.7	25.9	31.1	5.2	0.005	1.27
1BL8	3.2	28.0	29.0	1.0	32.5	32.7	0.2	0.006	1.10
1DP0	1.7	15.7	21.1	5.4	16.1	20.3	4.2	0.018	2.92
1SFC	2.4	26.5	30.3	3.8	28.2	31.2	3.1	0.009	1.30
2FNP	2.6	28.5	30.8	2.3	27.2	29.0	1.8	0.010	1.90
2QUK	2.8	26.7	29.0	2.3	25.8	27.3	1.4	0.009	1.60
2R4R	3.4	21.7	27.0	5.3	23.4	28.0	4.6	0.007	1.40
3CRW	4.0	23.7	31.9	8.2	21.8	30.0	8.2	0.008	4.40
3FFN	3.0	22.2	27.1	4.9	21.6	26.6	5.0	0.012	1.40
3HN8	4.5	23.8	25.9	2.1	25.2	27.4	2.3	0.004	0.94
Mean	3.1	24.1	28.3	4.2	24.8	28.4	3.6	0.009	1.82

[a] Including their resolution, R, R_{free}, $R_{free} - R$, and the RMS deviation (RMSD) of their bonds and angles from equilibrium values. Modest differences between the reported and re-calculated R, R_{free}, $R_{free} - R$ are expected. Except for this table, relative differences in R, R_{free}, $R_{free} - R$ due to the local optimization protocols are reported self-consistently using values calculated with FFX .

biomolecular crystals is reduced to 7.7 s on average for the AMOEBA potential using a desktop workstation, as shown in Table 3. For comparison, the same table is presented in Section B of the Supporting Information for the *direct* approximation to the AMOEBA potential, which reduces the average time to only 1.2 s. This illustrates that the most expensive part of the AMOEBA energy evaluation is the SCF rather than the permanent multipole electrostatics. For comparison, evaluation of the X-ray term of eq 1 for 1DP0 on 8 CPU cores costs a factor of 3 (∼50 s) more than the force field term (15 s on 8 CPUs + GPU). Since the AMOEBA polarizable force field is generally less expensive or about equal to the X-ray term, the computational cost between the X-ray and force field terms is balanced.

A subset of 10 crystallography data sets, described in detail in Table 4, were selected from the 30 used for timings to compare the quality of biomolecular X-ray rerefinement with and without AMOEBA electrostatics. It is known that rerefinement from deposited X-ray data with current methods including TLS treatment of B-factors in combination with a maximum likelihood X-ray target function improves most PDB entries.[69] We chose a broad resolution range, 1.7−4.5 Å with a mean of 3.1 Å, to promote more general conclusions. The averages of the originally reported R, R_{free}, and $R_{free} - R$ are 24.1, 28.3, and 4.2%, respectively. The averages of R, R_{free}, and $R_{free} - R$ recalculated using FFX gives 24.8, 28.4, and 3.6%, respectively, which provides a basis for self-consistent comparisons. Modest differences between the reported and recalculated R values are expected due to variations in the scattering engines of the various original refinement programs and FFX,[21] such as their treatment of bulk solvent and crystal anisotropy. All R values in Table 5 were calculated in FFX.

Beginning with the ten PDB structures listed in Table 4, a stringent local optimization procedure was applied using the refinement target given by eq 1 where the X-ray refinement term

Table 5. Refinement Statistics and MolProbity Metrics for 10 Biomolecular Crystallography Data Sets[a]

						poor	Ramachandran (%)		MolProbity
model	structure	R	R_{free}	$R_{free} - R$	clashscore	rot. (%)	outliers	favored	score
1A7B	PDB	25.9	31.1	5.2	24.9	8.3	0.8	95.1	2.93
3.1	vdW	20.5	31.6	11.1	1.4	14.1	0.3	92.6	2.20
	AMOEBA	20.6	30.9	10.3	4.4	12.2	0.0	97.0	1.86
1BL8	PDB	32.5	32.7	0.2	80.9	23.1	4.2	72.9	4.23
3.2	vdW	24.7	30.4	5.8	1.4	11.4	4.2	81.3	2.39
	AMOEBA	24.7	29.2	4.5	4.6	7.9	3.2	90.5	2.08
1DP0	PDB	16.1	20.3	4.2	10.6	3.6	0.2	96.5	2.20
1.7	vdW	14.7	18.9	4.2	3.7	2.1	0.1	97.3	1.53
	AMOEBA	14.9	19.0	4.1	3.8	1.9	0.1	97.3	1.51
1SFC	PDB	28.2	31.2	3.1	48.5	7.9	1.0	95.1	3.18
2.4	vdW	22.5	30.6	8.1	3.1	10.0	1.1	94.5	2.23
	AMOEBA	23.1	30.6	7.6	5.0	7.0	0.5	97.2	1.90
2FNP	PDB	27.2	29.0	1.8	75.0	9.8	5.8	82.5	3.80
2.6	vdW	21.0	30.4	9.4	2.4	15.4	2.5	87.1	2.54
	AMOEBA	21.1	28.3	7.2	5.5	13.3	2.1	92.5	2.34
2QUK	PDB	25.8	27.3	1.4	43.8	13.9	3.5	88.4	3.58
2.8	vdW	21.9	30.1	8.2	6.5	12.7	2.7	86.8	2.82
	AMOEBA	22.5	30.1	7.6	8.8	13.9	2.4	90.6	2.76
2R4R	PDB	23.4	28.0	4.6	80.3	11.3	4.4	79.1	3.92
3.4	vdW	20.5	27.1	6.6	4.3	13.6	4.6	81.3	2.79
	AMOEBA	20.9	26.7	5.8	7.1	12.5	2.4	87.3	2.66
3CRW	PDB	21.8	30.0	8.2	70.8	9.1	4.8	76.8	3.82
4.0	vdW	18.1	30.2	12.1	0.4	9.3	2.1	84.8	2.03
	AMOEBA	19.9	29.8	9.9	1.0	8.7	1.3	89.9	1.90
3FFN	PDB	21.6	26.6	5.0	13.1	10.9	1.6	94.2	2.81
3.0	vdW	16.9	25.2	8.3	1.8	11.2	1.4	92.9	2.19
	AMOEBA	17.1	24.5	7.4	3.4	10.1	1.5	95.6	2.01
3HN8	PDB	25.2	27.4	2.3	32.5	9.3	1.9	87.6	3.34
3.5	vdW	19.3	24.7	5.5	5.5	19.0	2.2	85.9	2.91
	AMOEBA	19.5	24.9	5.4	7.2	16.3	2.5	89.3	2.78
mean	**PDB**	24.8	28.4	3.6	48.0	10.7	2.8	86.8	3.38
	vdW	20.0	27.9	7.9	3.0	11.9	2.1	88.5	2.36
	AMOEBA	20.4	27.4	7.0	5.1	10.4	1.6	92.7	2.18

[a] R and R_{free} for the starting models (PDB) were re-calculated using FFX to enable self-consistent comparisons. Local optimizations were performed, as described in the text, without electrostatics (vdW) and using the AMOEBA polarizable force field (AMOEBA) for the chemical term of eq 1. The mean improvements in R_{free} are 0.5 and 1.0% under vdW and AMOEBA protocols, respectively. AMOEBA shows the greatest reduction in poor side-chain rotamers and Ramachandran backbone outliers (outliers). In addition, AMOEBA achieves the greatest increase in favorable backbone (θ,ϕ) dihedral pairs (favored) and overall MolProbity score. Although the vdW protocol achieves a lower clashscore than AMOEBA, this is due to incorrect treatment of weak hydrogen bonds (C−H···O) by this heuristic.

was described by Fenn et al.[21] and the chemistry term is a simplification of the AMOEBA potential chosen to mimic the REPEL force field used in CNS.[10] Specifically, both electrostatics and torsional terms were turned off to give a nonbonded force that only included van der Waals interactions (this potential is referred to by the abbreviation vdW below). The X-ray weight (w_a) was set to 2.5 based on optimization of the mean R_{free} for the 10 data sets following 10 rounds of coordinate and B-factor optimization under the vdW potential (data not shown), although a weight in the range of 1.0−5.0 does not change our conclusions. The coordinate optimizations during each round were converged to a RMS gradient of 0.05 kcal/mol/Å, and the B-factor optimizations were converged to a RMS gradient of 0.005 (unitless). Beginning from the final vdW model, further optimization was performed using eq 1 based on either the full AMOEBA force field (referred to as AMOEBA) or the *direct* polarization approximation to AMOEBA (results in Section C of the Supporting Information).

The final models from the vdW and AMOEBA local optimization protocols will first be compared based on R_{free}, $R_{free} - R$, and local structural metrics computed using MolProbity,[70,71] as shown in Table 5. The AMOEBA optimization reduced R_{free} by an average of 0.5% relative to the starting models obtained via the vdW optimization procedure. In addition, inclusion of electrostatics in the AMOEBA optimizations reduced overfitting ($R_{free} - R$) by 0.9%. It is important to emphasize that although the R, R_{free} and $R_{free} - R$ reported in Table 5 are calculated self-consistently in FFX, comparisons between the deposited

Table 6. Geometric Statistics, Coordinate Superposition RMSDs and the Relative Energy per Residue for 10 Biomolecular Crystallography Data Sets[a]

model		geometry RMSD		coord. RMSD (Å)		
res (Å)	potential	bond (Å)	angle (°)	C_α	heavy	rel. energy/residue (kcal/mol)
1A7B	vdW	0.014	2.64	0.39	0.67	0.0
3.1	AMOEBA	0.013	2.77	0.43	0.77	−20.0
1BL8	vdW	0.017	3.08	0.60	0.87	0.0
3.2	AMOEBA	0.016	3.27	0.74	0.98	−7.3
1DP0	vdW	0.020	2.94	0.11	0.24	0.0
1.7	AMOEBA	0.020	3.07	0.11	0.25	−16.4
1SFC	vdW	0.014	2.74	0.35	0.62	0.0
2.4	AMOEBA	0.014	3.07	0.42	0.75	−9.1
2FNP	vdW	0.016	2.76	0.44	0.82	0.0
2.6	AMOEBA	0.016	3.06	0.50	0.96	−9.9
2QUK	vdW	0.015	2.79	0.30	0.56	0.0
2.8	AMOEBA	0.015	3.04	0.30	0.59	−6.8
2R4R	vdW	0.015	2.65	0.60	0.90	0.0
3.4	AMOEBA	0.014	2.94	0.64	0.98	−6.7
3CRW	vdW	0.014	2.39	0.59	0.92	0.0
4.0	AMOEBA	0.014	2.83	0.92	1.24	−12.0
3FFN	vdW	0.015	2.68	0.31	0.48	0.0
3.0	AMOEBA	0.014	2.95	0.31	0.52	−7.0
3HN8	vdW	0.016	2.82	0.44	0.68	0.0
3.5	AMOEBA	0.016	3.12	0.45	0.73	−8.4
mean	**vdW**	0.016	2.75	0.41	0.67	0.0
	AMOEBA	0.015	3.01	0.48	0.78	−10.4

[a] The structures used here correspond to those of Table . Compared to the vdW structures, inclusion of electrostatics slightly decreased the bond RMSD from equilibrium but increased the angle RMSD. The Cα coordinate RMSD, computed relative to the starting PDB structures, was 0.41 and 0.48 Å under the vdW and AMOEBA protocols, respectively. The mean heavy atom coordinate RMSD was 0.67 and 0.78 Å for vdW and AMOEBA, respectively. Finally, the AMOEBA potential energy per residue, relative to the vdW structure, was reduced by 10.4 kcal/mol by optimization with AMOEBA polarizable electrostatics.

structures and vdW minima are not significant it terms of drawing conclusions with respect to the merits of a potential energy function. What is significant, however, is the reduction in R_{free} and overfitting upon local optimization from the baseline vdW minima using the full AMOEBA model, with all other adjustable parameters fixed. We also note the increase in the average $R_{free} - R$ in going from the deposited PDB structures to the vdW minima. This is explained by the original structures being refined without hydrogens and/or not being optimized to a tight convergence criterion.

The vdW optimization drastically reduced the van der Waals clashscore from a mean of 48.0 to only 3.0, which is the number of van der Waals clashes per 1000 atoms. Similar improvements can be achieved via the all-hydrogen force field in the initial[37] and more recent[10] versions of CNS. We note that formation of energetically favorable weak hydrogen bonds (i.e., C−H···O) that are driven by electrostatics are incorrectly counted as clashes by MolProbity (among other simplifications). This explains why the AMOEBA protocol causes a modest increase in clashscore relative to the vdW potential and points out a limitation of the generally useful MolProbity clashscore heuristic.

Although the percentage of poor side-chain rotamers was increased by the vdW optimization from a mean of 10.7 to 11.9%, the backbone Ramachandran statistics improved. Specifically, the percentage of outliers was reduced from 2.8 to 2.1% and the percentage of favorable (ϕ,φ) dihedral pairs increased from 86.8 to 88.5%. Unlike the vdW result, inclusion of electrostatics in the AMOEBA protocol slightly reduced the percentage of poor side-rotamers to 10.4%. Backbone Ramachandran outliers were further reduced from the vdW result to 1.6% under AMOEBA, and favorable (ϕ,φ) dihedral pairs were further improved from the vdW result by 4.2%.

The overall MolProbity score is a log-weighted combination of the clashscore, percentage of bad side-chain rotamers, and the unfavored Ramachandran percentage that indicates the crystallographic resolution at which those values would be expected.[71] The mean value of the starting models is 3.38, which is slightly worse than the actual average crystallographic resolution of 3.1 Å. Under the vdW and AMOEBA protocols, the score was improved to 2.36 and 2.18, respectively. Therefore, MolProbity judges the quality of the AMOEBA structures to be consistent with a mean crystallographic resolution that is 0.92 Å better than the actual mean of the 10 data sets. Without going into details, the contribution from the clashscore was fixed to the vdW result when calculating MolProbity scores for AMOEBA (and direct) to ameliorate limitations of the clashscore heuristic for weak hydrogen bonds.

The RMS deviation of the bonds and angles from equilibrium values for the optimized structures referred to in Table 5 are listed in Table 6. Relative to the starting models, the bond RMSD increased from 0.009 to 0.015 Å and the angle RMSD from 1.822 to 2.749°. The increase of approximately 50% in both cases may

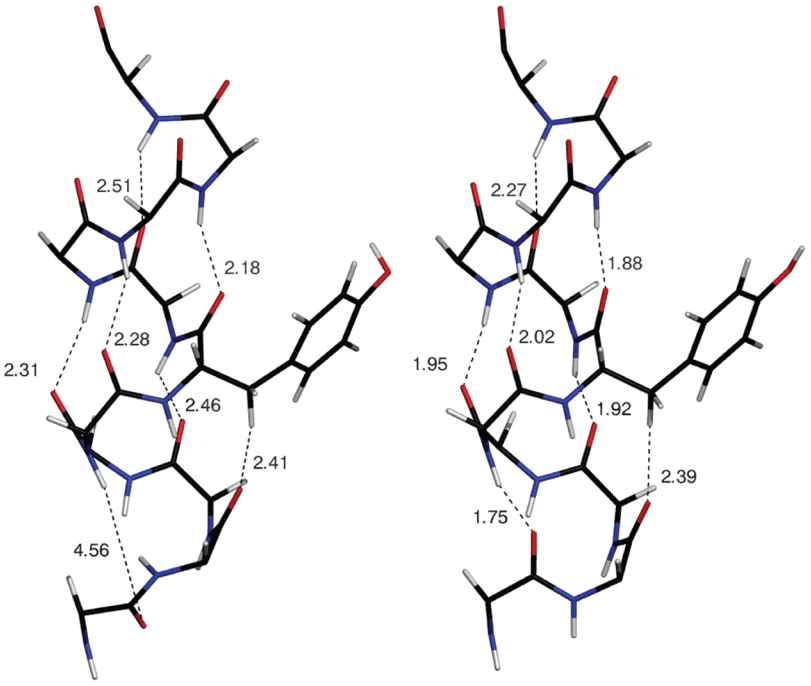

Figure 5. Panels A and B show residues A214−A224 of the human β_2 adrenergic G-protein-coupled receptor (2R4R) after optimization with the vdW and AMOEBA protocols, respectively. Canonical i to $i + 4$ α-helix hydrogen-bonding distances are explicitly drawn. The vdW protocol (Panel A) exhibits five potential canonical hydrogen bonds, but the mean N−H···O distance (2.35 Å) is relatively large. The AMOEBA protocol (Panel B) lengthened the helix by one residue relative to the vdW protocol. These six α-helix hydrogen bonds show a mean N−H···O distance (1.97 Å) in the optimal energetic range. We note that the second AMOEBA oxygen from the bottom of the image forms a weak hydrogen bond to the $i + 4$ C_β−H instead of to the $i + 4$ N−H, which could be improved by replacing the local optimization protocol used in this work with a global one.

be explained by the relatively stiff bond and angle force constants often used by crystallographic refinement software compared to AMOEBA force constants that are fit to vacuum vibrational frequencies measured experimentally or determined by electronic structure calculations. Alternatively, the increase could be due to differences in the weighting (w_a) between the X-ray and chemical terms in the target function. The mean energy stored in the bonds and angles of the final AMOEBA structures was only 0.08 and 0.14 kcal/mol, respectively, which is easily less than kT at the range of temperatures crystallographic data is collected (100−300 K).

The RMS deviation in the atomic coordinates for C_α atoms and for all heavy atoms from the PDB starting structures is also shown in Table 6. The vdW and AMOEBA protocols moved C_α atoms by an average of 0.41 and 0.49 Å, respectively. Larger RMS coordinate deviations of 0.67 and 0.78 Å were observed for all heavy atoms after the vdW and AMOEBA protocols, respectively. A relative potential energy per residue was calculated by subtracting the asymmetric unit potential energy of the AMOEBA model from that of the vdW model using the full AMOEBA potential energy function (AMOEBA is a better estimate of the true potential energy than the vdW potential) followed by division by the number of residues to achieve an estimate independent of protein size. The relatively small local perturbation of the coordinates due to optimization with electrostatics nonetheless resulted in a lowering of the relative potential energy per residue by more than 10 kcal/mol, which is comparable to a protein/drug binding free energy (for example, benzamidine binding to trypsin is favorable by 6.3−7.3 kcal/mol).[7] The dramatic energetic improvement is consistent with electrostatic stabilization from the formation of hydrogen bonds, as shown in Figure 5 for an α-helix of the human β_2 adrenergic G-protein-coupled receptor (2R4R, 3.4 Å). The AMOEBA protocol lengthened the α-helix by one residue relative to the vdW protocol, which is consistent with the 1.0 Å higher resolution 2RH1 structure (2RH1, 2.4 Å, Figure S-1 in the Supporting Information).

V. CONCLUSIONS

From this work we conclude the AMOEBA polarizable force field evaluated with PME electrostatics is capable of improving macromolecular X-ray crystallography refinement starting from structures obtained by optimization with only van der Waals nonbonded forces. The improvements lower R_{free} by 0.5%, reduce overfitting by 0.9% and increase the number of residues with favorable backbone conformations by 4.2%. This is consistent with electrostatics driving local conformational shifts toward favorable hydrogen-bonding networks, especially for repetitive secondary structure, as shown in Figure 5. The mean MolProbity score for our final models suggests geometric quality consistent with a mean crystallographic resolution of 2.18 Å, which is 0.92 Å better than the true mean of 3.1 Å.

We have shown that PME electrostatics benefits from the explicit incorporation of space group symmetry to reduce both memory and CPU demands. Further acceleration was achieved using shared memory parallelization and a GPU coprocessor for the $N \cdot \log(N)$ reciprocal space convolution of the PME algorithm. The X-ray crystallography refinements were carried out in FFX, which currently depends on v. 1.6 of the JRE and v. 3.1 of

the CUDA API for additional acceleration using a GPU coprocessor. Relative to a single CPU core after expansion to P1, the combination of space group symmetry, shared memory parallelization over 8 Intel Xeon E5530 CPU cores at 2.4 GHz, and a Tesla M1060 GPU coprocessor at 1.30 GHz showed an average speed-up of more than 24× for large marcomolecular crystals that average 240 000 atoms in the unit cell.

One limitation of our results is the lack of a physical treatment of bulk solvent, such as Poisson−Boltzmann[72] (PB) or generalized Kirkwood[73] (GK) continuum electrostatics. Currently, the PB and GK models for AMOEBA do not include explicit support for symmetry operators or periodic boundary conditions, but it should be possible to extend them in this respect. Although the cost of numerical solutions to the PB equation for AMOEBA is prohibitive for macromolecular X-ray refinement, it may be possible to combine the analytic GK approximation with PME. It has been noted previously that global optimization via simulated annealing may lead to unreasonable side-chain conformations without continuum solvent, especially for charged residues at the surface of a macromolecule that incorrectly experience a vacuum environment.[32] Although the addition of a continuum solvent[33] followed by global optimization via simulated annealing[68] has been suggested, the fundamental problem of how to combine analytic continuum electrostatics with a rigorous lattice summation method remains an open question.

Although it is beyond the scope of this work, it is of interest to compare the improvements in model quality from AMOEBA electrostatics evaluated with PME to refinement using fixed charge electrostatics evaluated with spherical cutoffs as in CNS[10] or to electronic structure methods.[74,75] For example, the general features of the hydrogen-bonding network in Figure 5 might be reproduced by refinement with a fixed charge force field. However, the advantages of the polarizable AMOEBA model over fixed charge potentials have already been studied in detail for the structural properties of water,[5,6] ion solvation thermodynamics,[76] protein−ligand binding affinities,[7,77] and small molecule structural and thermodynamic observables.[8] Similar advantages have also been observed for the CHARMM polarizable force field based on the classical drude oscillator.[78−81]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. The definition of the AMOEBA *direct* polarization approximation and self-consistent field procedure. Timings and refinements for the test systems using direct polarization are then presented. Analysis of single precision and double precision force accuracy in the context of macromolecular X-ray refinement. One additional figure is presented that demonstrates AMOEBA-assisted refinement of 2R4R (3.4 Å). This information is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: michael.schnieders@gmail.com. Telephone: (650) 995-3526.

■ ACKNOWLEDGMENT

The authors acknowledge Pengyu Ren, Thomas A. Darden, Alan M. Grossfield, and Jay W. Ponder for the PME code in TINKER this work began from and for helpful discussions. The authors also wish to thank Axel T. Brunger for suggestions with regards to formulating self-consistent tests of refinement target functions. This work has been supported by an award from the NSF to Vijay S. Pande, Jay W. Ponder, Teresa Head-Gordon, and Martin Head-Gordon for "Collaborative Research: Cyberinfrastructure for Next Generation Biomolecular Modeling" (award number CHE-0535675) and by the Howard Hughes Medical Institute. For computer resources we acknowledge NSF award CNS-0619926 that supports the Bio-X2 cluster.

■ REFERENCES

(1) Schnieders, M. J.; Fenn, T. D.; Pande, V. S.; Brunger, A. T. Polarizable atomic multipole X-ray refinement: application to peptide crystals. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65* (9), 952–965.

(2) Fenn, T. D.; Schnieders, M. J.; Brunger, A. T.; Pande, V. S. Polarizable atomic multipole X-ray refinement: hydration geometry and application to macromolecules. *Biophys. J.* **2010**, *98* (12), 2984–2992.

(3) Fenn, T. D.; Schnieders, M. J.; Mustyakimov, M.; Wu, C.; Langan, P.; Pande, V. S.; Brunger, A. T. Reintroducing electrostatics into macromolecular crystallographic refinement: application to neutron crystallography and DNA hydration. *Structure* **2011**, 19.

(4) Ren, P.; Ponder, J. W. Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J. Comput. Chem.* **2002**, *23* (16), 1497–1506.

(5) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107* (24), 5933–5947.

(6) Ren, P.; Ponder, J. W. Temperature and pressure dependence of the AMOEBA water model. *J. Phys. Chem. B* **2004**, *108* (35), 13427–13437.

(7) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. Calculation of protein-ligand binding free energy by using a polarizable potential. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (17), 6290–6295.

(8) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(9) Ponder, Jay W. *TINKER: Software Tools for Molecular Design*, 5.0; Jay W. Ponder: Saint Louis, MO, 2009.

(10) Brunger, A. T., Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2007**, *2*, (11), 2728-2733.

(11) Darden, T.; York, D.; Pedersen, L. Particle-mesh Ewald - An n log(n) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(12) Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **1921**, *369* (3), 253–287.

(13) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle-mesh Ewald method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

(14) Sagui, C.; Darden, T. A. Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155–179.

(15) Toukmaji, A.; Sagui, C.; Board, J.; Darden, T. Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* **2000**, *113* (24), 10913–10927.

(16) Sagui, C.; Pedersen, L. G.; Darden, T. A. Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **2004**, *120* (1), 73–87.

(17) Shan, Y. B.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* **2005**, *122* (5), 13.

(18) Cerutti, D. S.; Case, D. A. Multi-level Ewald: A hybrid multigrid/fast Fourier transform approach to the electrostatic particle-mesh problem. *J. Chem. Theory Comput.* **2010**, *6* (2), 443–458.

(19) Cerutti, D. S.; Duke, R. E.; Darden, T. A.; Lybrand, T. P. Staggered mesh ewald: an extension of the smooth particle-mesh Ewald method adding great versatility. *J. Chem. Theory Comput.* **2009**, *5* (9), 2322–2338.

(20) Neelov, A.; Holm, C. Interlaced P3M algorithm with analytical and ik-differentiation. *J. Chem. Phys.* **2010**, *132* (23), 15.

(21) Fenn, T. D.; Schnieders, M. J.; Brunger, A. T. A smooth and differentiable bulk-solvent model for macromolecular diffraction. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66* (9), 1024–1031.

(22) Karttunen, M.; Rottler, J.; Vattulainen, I.; Sagui, C. Electrostatics in biomolecular simulations: Where are we now and where are we heading? In *Computational Modeling of Membrane Bilayers*; Elsevier Academic Press Inc.: San Diego, CA, 2008; Vol. 60, pp 49−89.

(23) Stone, A. J.; Alderton, M. Distributed multipole analysis-methods and applications. *Mol. Phys.* **1985**, *56* (5), 1047–1064.

(24) Stone, A. J. Intermolecular potentials. *Science* **2008**, *321* (5890), 787–789.

(25) Ponder, J. W.; Case, D. A. Force fields for protein simulations. In *Advances in Protein Chemistry*; Academic Press: San Diego, CA, 2003; Vol. 66, pp 27−85.

(26) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D.; Roux, B. Atomic level anisotropy in the electrostatic modeling of lone pairs for a polarizable force field based on the classical Drude oscillator. *J. Chem. Theory Comput.* **2006**, *2* (6), 1587–1597.

(27) Rafat, M.; Popelier, P. L. A. A convergent multipole expansion for 1,3 and 1,4 Coulomb interactions. *J. Chem. Phys.* **2006**, *124* (14), 7.

(28) Rafat, M.; Shaik, M.; Popelier, P. L. A. Transferability of quantum topological atoms in terms of electrostatic interaction energy. *J. Phys. Chem. A* **2006**, *110* (50), 13578–13583.

(29) Solano, C. J. F.; Pendas, A. M.; Francisco, E.; Blanco, M. A.; Popelier, P. L. A. Convergence of the multipole expansion for 1,2 Coulomb interactions: The modified multipole shifting algorithm. *J. Chem. Phys.* **2010**, *132* (19), 10.

(30) Afonine, P. V.; Grosse-Kunstleve, R. W.; Adams, P. D.; Lunin, V. Y.; Urzhumtsev, A. On macromolecular refinement at subatomic resolution with interatomic scatterers. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *63*, 1194–1197.

(31) Dawson, B., The covalent bond in diamond. *Proc. R. Soc. London, Ser. A* **1967**, *298*, (1454), 264−288.

(32) Weis, W. I.; Brunger, A. T.; Skehel, J. J.; Wiley, D. C. Refinement of the influenza-virus hemagglutinin by simulated annealing. *J. Mol. Biol.* **1990**, *212* (4), 737–761.

(33) Moulinier, L.; Case, D. A.; Simonson, T. Reintroducing electrostatics into protein X-ray structure refinement: bulk solvent treated as a dielectric continuum. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 2094–2103.

(34) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101* (16), 3005–3014.

(35) Bashford, D.; Case, D. A. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

(36) 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.

(37) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J.-S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR System: A new software suite for macromolecular structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 905–921.

(38) Adams, P. D.; Grosse-Kunstleve, R. W.; Hung, L. W.; Ioerger, T. R.; McCoy, A. J.; Moriarty, N. W.; Read, R. J.; Sacchettini, J. C.; Sauter, N. K.; Terwilliger, T. C. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 1948–1954.

(39) Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(40) Bricogne, G.; Blanc, E.; Brandl, M.; Flensburg, C.; Keller, P.; Paciorek, P.; Roversi, P.; Sharff, A.; Smart, O.; Vonrhein, C.; Womack, T. *BUSTER*, 2.9; Global Phasing Ltd.: Cambridge, U.K., 2010.

(41) Bailey, S., The CCP4 suite - programs for protein crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50*, 760-763.

(42) Smith, W., The minimum image convention in non-cubic MD cells. *CCP5 Information Quaterly* **1989**, *30*, (35).

(43) Smith, E. R. Electrostatic energy in ionic crystals. *Proc. R. Soc. London, Ser. A* **1981**, *375* (1763), 475–505.

(44) Deleeuw, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. 1. Lattice sums and dielectric constants. *Proc. R. Soc. London, Ser. A* **1980**, *373* (1752), 27–56.

(45) Deleeuw, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. 2. Equivalence of boundary conditions. *Proc. R. Soc. London, Ser. A* **1980**, *373* (1752), 57–66.

(46) Deleeuw, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. 3. Further theory and applications. *Proc. R. Soc. London, Ser. A* **1983**, *388* (1794), 177–193.

(47) Stone, A. J. *The Theory of Intermolecular Forces.*: Clarendon Press: Oxford, 1996; Vol. 32, p 264.

(48) Stone, A. J. Distributed multipole analysis: Stability for large basis sets. *J. Chem. Theory Comput.* **2005**, *1* (6), 1128–1132.

(49) Shi, Y.; Wu, C.; Ponder, J. W.; Ren, P. Multipole Electrostatics in Hydration Free Energy Calculations. *J. Chem. Comput.* **2010**, 32.

(50) Smith, W. Point multipoles in the Ewald summation (revisited). *CCP5 Information Quaterly* **1982**, *4*, 13.

(51) Thole, B. T. Molecular polarizabilities calculated with a modified dipole interaction. *Chem. Phys.* **1981**, *59* (3), 341–350.

(52) McMurchie, L. E.; Davidson, E. R. One- and two-electron integrals over Cartesian Gaussian functions. *J. Comput. Phys.* **1978**, *26* (2), 218–231.

(53) Challacombe, M.; Schwegler, E.; Almlof, J. Recurrence relations for calculation of the Cartesian multipole tensor. *Chem. Phys. Lett.* **1995**, *241* (1−2), 67–72.

(54) Brunger, A. A memory-efficient fast Fourier transformation algorithm for crystallographic refinement on supercomputers. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1989**, *45*, 42–50.

(55) Aguado, A.; Madden, P. A. Ewald summation of electrostatic multipole interactions up to the quadrupolar level. *J. Chem. Phys.* **2003**, *119* (14), 7471–7483.

(56) Agarwal, R. C. New least-squares refinement technique based on fast Fourier-transform algorithm. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 791–809.

(57) Bowers, K. J.; Dror, R. O.; Shaw, D. E. Zonal methods for the parallel execution of range-limited N-body simulations. *J. Comput. Phys.* **2007**, *221* (1), 303–329.

(58) Frigo, M.; Johnson, S. G. The design and implementation of FFTW3. *Proc. IEEE* **2005**, *93* (2), 216–231.

(59) Neumann, M. A.; Leusen, F. J. J.; Kendrick, J. A major advance in crystal structure prediction. *Angew. Chem., Int. Ed.* **2008**, *47* (13), 2427–2430.

(60) Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J. S.; Della Valle, R. G.; Venuti, E.; Jose, J.; Gadre, S. R.; Desiraju, G. R.; Thakur, T. S.; van Eijck, B. P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Neumann, M. A.; Leusen, F. J. J.; Kendrick, J.; Price, S. L.; Misquitta, A. J.; Karamertzanis, P. G.; Welch, G. W. A.; Scheraga, H. A.; Arnautova,

Y. A.; Schmidt, M. U.; van de Streek, J.; Wolf, A. K.; Schweizer, B. Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2009**, *65*, 107–125.

(61) Perrin, M. A.; Neumann, M. A.; Elmaleh, H.; Zaske, L. Crystal structure determination of the elusive paracetamol Form III. *Chem. Commun.* **2009**, *22*, 3181–3183.

(62) Davies, J. M.; Brunger, A. T.; Weis, W. I. Improved structures of full-length p97, an AAA ATPase: Implications for mechanisms of nucleotide-dependent conformational change. *Structure* **2008**, *16* (5), 715–726.

(63) Brunger, A. T.; DeLaBarre, B.; Davies, J. M.; Weis, W. I. X-ray structure determination at low resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 128–133.

(64) Bazterra, V. E.; Thorley, M.; Ferraro, M. B.; Facelli, J. C. A distributed computing method for crystal structure prediction of flexible molecules: An application to N-(2-dimethyl-4,5-dinitrophenyl) acetamide. *J. Chem. Theory Comput.* **2007**, *3* (1), 201–209.

(65) Selmer, M.; Dunham, C. M.; Murphy, F. V.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **2006**, *313* (5795), 1935–1942.

(66) Robertus, J. D.; Ladner, J. E.; Finch, J. T.; Rhodes, D.; Brown, R. S.; Clark, B. F. C.; Klug, A. Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **1974**, *250* (5467), 546–551.

(67) Shi, H. J.; Moore, P. B. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA* **2000**, *6* (8), 1091–1105.

(68) Brunger, A. T. Simulated annealing in crystallography. *Annu. Rev. Phys. Chem.* **1991**, *42*, 197–223.

(69) Joosten, R. P.; Womack, T.; Vriend, G.; Bricogne, G. Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 176–185.

(70) Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B.; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007**, *35*, W375–W383.

(71) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *66*, 12–21.

(72) Schnieders, M. J.; Baker, N. A.; Ren, P. Y.; Ponder, J. W. Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum. *J. Chem. Phys.* **2007**, *126*, 12.

(73) Schnieders, M. J.; Ponder, J. W. Polarizable atomic multipole solutes in a generalized Kirkwood continuum. *J. Chem. Theory Comput.* **2007**, *3* (6), 2083–2097.

(74) Yu, N.; Li, X.; Cui, G. L.; Hayik, S. A.; Merz, K. M. Critical assessment of quantum mechanics based energy restraints in protein crystal structure refinement. *Protein Sci.* **2006**, *15* (12), 2773–2784.

(75) Li, X.; Hayik, S. A.; Merz, K. M. QM/MM X-ray refinement of zinc metalloenzymes. *J. Inorg. Biochem.* **2010**, *104*, 512–522.

(76) Grossfield, A.; Ren, P. Y.; Ponder, J. W. Ion solvation thermodynamics from simulation with a polarizable force field. *J. Am. Chem. Soc.* **2003**, *125* (50), 15671–15682.

(77) Jiao, D.; Zhang, J. J.; Duke, R. E.; Li, G. H.; Schnieders, M. J.; Ren, P. Y. Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential. *J. Comput. Chem.* **2009**, *30* (11), 1701–1711.

(78) Lopes, P. E. M.; Roux, B.; MacKerell, A. D. Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications. *Theor. Chem. Acc.* **2009**, *124* (1−2), 11–28.

(79) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D. Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. *J. Phys. Chem. B* **2007**, *111* (11), 2873–2885.

(80) Lamoureux, G.; Roux, B. Absolute hydration free energy scale for alkali and halide ions established from simulations with a polarizable force field. *J. Phys. Chem. B* **2006**, *110* (7), 3308–3322.

(81) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* **2006**, *418* (1−3), 245–249.