Force Field X: A Computational Microscope to Study Genetic Variation and Organic Crystals Using Theory and Experiment

Rose A Gogal, Aaron J Nessler, Andrew C Thiel, Hernan V Bernabe, Rae A Corrigan Grove, Leah M Cousineau, Jacob M Litman, Jacob M Miller, Guowei Qi, Matthew J Speranza, et al.

The Journal of Chemical Physics

AIP Publishing

Force Field X: A Computational Microscope to Study Genetic Variation and Organic Crystals Using Theory and Experiment

Rose A. Gogal[1,#], Aaron J. Nessler[1,#], Andrew C. Thiel[1,#], Hernan V. Bernabe[1], Rae A. Corrigan Grove[3], Leah M. Cousineau[2], Jacob M. Litman[2], Jacob M. Miller[1], Guowei Qi[2], Matthew J. Speranza[1], Mallory R. Tollefson[1], Timothy D. Fenn[4], Jacob J. Michaelson[5], Okimasa Okada[6], Jean-Philip Piquemal[7], Jay W. Ponder[8], Jana Shen[9], Richard J. H. Smith[10], Wei Yang[11,12], Pengyu Ren[13], Michael J. Schnieders[1,2,*]

[1]Roy J. Carver Department of Biomedical Engineering, University of Iowa, Iowa City, IA, 52242, USA

[2]Department of Biochemistry and Molecular Biology, University of Iowa, Iowa City, IA, 52242, USA

[3]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[4]Analytical Development, LEXEO Therapeutics, New York, NY 10010, USA

[5]Department of Psychiatry, University of Iowa Hospitals and Clinics, Iowa City, IA, 52242, USA

[6]Sohyaku Innovative Research Division, Mitsubishi Tanabe Pharma Corporation, 1000 Kamoshida-cho, Aoba-ku, Yokohama, Kanagawa 227-0033, Japan

[7]Department of Chemistry, Sorbonne Université, F-75005 Paris, France

[8]Department of Chemistry, Washington University in St. Louis, St. Louis, MO, 63130, USA

[9]Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, MD, USA

[10]Molecular Otolaryngology & Renal Research Laboratories, Department of Otolaryngology, University of Iowa Hospitals and Clinics, Iowa City, IA, 52242, USA

[11]Department of Chemistry and Biochemistry, Florida State University, FL 32309, USA

[12]Institute of Molecular Biophysics, Florida State University, FL 32309, USA

[13]Department of Biomedical Engineering, University of Texas, Austin, TX, 78712, USA

[#]Denotes Joint First Authors

*Correspondence: michael-schnieders@uiowa.edu

Abstract

Force Field X (FFX) is an open-source software package for atomic resolution modeling of genetic variants and organic crystals that leverages advanced potential energy functions and experimental data. FFX currently consists of nine modular packages with novel algorithms that include global optimization via a many-body expansion, acid-base chemistry using polarizable constant-pH molecular dynamics, estimation of free energy differences, generalized Kirkwood implicit solvent models, and many more. Applications of FFX focus on use and development of a crystal structure prediction pipeline, biomolecular structure refinement against experimental datasets, and estimation of the thermodynamic effects of genetic variants on both proteins and nucleic acids. Use of Parallel Java and OpenMM combine to offer shared memory, message passing, and GPU parallelization for high performance simulations. Overall, the FFX platform serves as a computational microscope to study systems ranging from organic crystals to solvated biomolecular systems.

Introduction

Force Field X (FFX) is an open-source software platform for atomic resolution modeling of organic materials and biomolecules that leverages molecular mechanics force fields and a family of novel electrostatics[1-3], optimization[4,5], thermodynamics[6-8], and experimental refinement algorithms[9-11]. This article introduces the unique computational models and algorithms available in FFX, with emphasis placed on applications such as crystal structure prediction, understanding the mechanism of disease-causing protein missense variants, and the refinement of biomolecules against experimental datasets. FFX was envisioned as a platform to develop experimental refinement algorithms using advanced potential energy functions based on permanent atomic multipoles and induced dipoles, including the Atomic Multipole Optimized Energetics for Biomolecular Applications (AMOEBA) force field[12-16]. FFX's design has been influenced by lessons learned from packages such as Tinker[17-19], OpenMM[20] and CNS[21,22], while prioritizing the use of virtual machine (VM) technology and polyglot programming. FFX is distributed under the

GPL v.3 license with the classpath exception, which is the same license used by OpenJDK (although v.2 in this case). The long-term goal for the FFX platform is to support the study of all major facets of organic materials, biological molecules, and their assembly into large complexes using the principles of chemical physics.

Polyglot Architecture. FFX achieves portability and support for polyglot programming (*i.e.*, the use of several programming languages) through VM technologies, including the Java VM (JVM)[23-25], GraalVM[26,27], and TornadoVM[28,29] as shown in Figure 1. The JVM, as part of the Java Development Kit (JDK), is one of the oldest VM platforms. Over the last decade, the GraalVM project has augmented the JVM with an alternative just-in-time compiler (Graal), the ability to create ahead-of-time native images, and support for polyglot programming to allow languages such as Python[30,31] to execute on the JVM and interoperate with canonical JVM languages (*e.g.*, Java, Kotlin, Scala, and Groovy).

Figure 1. An overview of Force Field X. The second row shows languages that are compatible with FFX. The third row represents the platforms that execute FFX on CPU cores on the left and coprocessors on the right. The final row indicates the CPU and GPU hardware that FFX can perform calculations on.

Novel Algorithms. The platform includes many novel algorithms that will be discussed in more detail throughout this article. For example, FFX debuted the first implementation of X-ray

crystallography refinement using the polarizable atomic multipole AMOEBA force field, the first constant-pH molecular dynamics support for a polarizable force field[32], and the generalized Kirkwood (GK) implicit solvent models[3,33-35]. FFX also implements efficient inclusion of space group symmetry into long range electrostatics via particle-mesh Ewald summation[1,36-38] and orthogonal space tempering for the calculation of free energy differences[6,8,39,40]. It can also perform global optimization using dead-end[41] (and Goldstein[42]) elimination for target functions that include higher order many-body terms[11,43] and dual topology methods to compute free energy differences between potential energy models[7,44].

Commands & Parallelization. FFX offers more than 50 commands that implement a variety of methods to investigate organic and biomolecular systems, which are organized into nine Java packages. These packages will be discussed in more detail later in this article. FFX leverages GraalVM technologies for cross-platform polyglot programming in Java, Groovy, Python, and Kotlin. FFX utilizes the Parallel Java (PJ) package to facilitate parallelization across the CPU cores/threads of a single process (*i.e.*, shared memory) and among multiple processes using its message passing interface (MPI). GPU acceleration is currently achieved using OpenMM, based on a Java class hierarchy that mirrors the OpenMM C++ API. The source code is freely available for download from the FFX website (http://ffx.biochem.uiowa.edu) or from GitHub (https://github.com/SchniedersLab/forcefieldx) and currently depends on JDK version 21 or greater. The software is compiled within a terminal window using the Apache Maven tool and is executed in either "headless" command line mode using the "ffxc" command or with a simple interactive graphical user interface using the "ffx" command.

Applications. Throughout this article, we will discuss applications of the novel algorithms available in FFX. One application is the refinement of atomic resolution models against X-ray and/or neutron diffraction data using advanced force fields[10,11]. FFX is also able to investigate the impact of missense variants on protein structure, dynamics, and thermodynamics[45-56]. In addition,

FFX premiers a novel pipeline for crystal structure prediction[57,58]. The wide-ranging capabilities of FFX make it particularly useful for the computational study of biochemical mechanisms.

Features and Organization

File Types, Coordinate Representations, and Conversions. FFX utilizes a variety of file types to describe molecular systems. Files typically consist of a base name that is shared among all files for a system and a unique suffix to denote the information contained therein (*e.g.*, molecule.xyz, molecule.mtz, molecule.properties, *etc.*). FFX uses a suffix versioning system where subsequent files of the same type are saved with an appended integer (*e.g.*, generated coordinates from molecule.xyz would produce a file named molecule.xyz_2). Commonly used file types are found with a brief description in Table *1*.

Table 1. File types read and written by FFX, including the origin of each format.

Atomic Coordinates and Variables		
ARC	Structural archive	Tinker
CIF	Crystallographic Information File	IUCr
DYN	Molecular dynamics restart information	Tinker
ESV	Extended system variables	FFX
INT	Internal coordinates	Tinker
PDB	Protein Databank structure	PDB
XPH	XYZ file with pH information	FFX
XYZ	Coordinates, atom types, and connectivity	Tinker
Control Properties and Force Field Parameters		
DST	Distance matrix from superposing structures	FFX
Properties/Key	Control file with Java properties/Tinker keywords	FFX/Tinker
PRM	Force field parameter file	Tinker
Structure Factors and Real Space Maps		
CCP4/MAP	Real space density map	CCP4
CIF	Structure factors	IUCr
CNS/HKL	Structure factors	CNS/Xplor
MTZ	Binary format for structure factors	CCP4
XPLOR	Real space density map	CNS/Xplor
Thermodynamics		
BAR	Window energy values for FEP/BAR	Tinker
HIS	Orthogonal space histogram	FFX
LAM	Lambda restart information	FFX
MBAR	Window energy values for MBAR	FFX

FFX implements a variety of commands to facilitate alterations and conversions between the file types listed in Table *1*. General file commands are listed in Table *2* along with a brief description. The *Cart2Frac* and *Frac2Cart* commands convert atomic coordinates between Cartesian and fractional coordinate systems, where the latter defines each atomic position as a fractional (unitless) distance along each axis of a unit cell. *SaveAsP1* expands a periodic system by applying space group symmetry operators to the asymmetric unit to generate a P1 unit cell (or a larger replicated unit cell). *SaveAsXYZ* converts a system into an XYZ coordinate file. *SaveAsQE* generates a default script that can be used with Quantum ESPRESSO to perform a plain-wave self-consistent field (PWscf) calculation. *ImportCIF* converts an organic crystal from CIF format (*e.g.*, from the Cambridge Structural Database[59]) into XYZ format based on the

constituent molecule(s) having already been parameterized. While reading systems, *ImportCIF* actively enforces space group restrictions with regards to lattice parameters and updates the space group in some cases (*e.g.*, a rhombohedral space group with hexagonal lattice parameters is converted to the appropriate hexagonal space group). *ImportCIF* adds hydrogen atoms to a system if none are present in the original CIF file. It can also convert an XYZ system to a basic CIF format containing the coordinate and crystal information. *MoveIntoUnitCell* moves the molecules of a system to ensure each center of mass is located within the unit cell. Finally, the command *xray.MTZInfo* displays human readable information for a binary MTZ file while *xray.CIFtoMTZ* generates an MTZ file for a corresponding CIF file.

Table 2. Structure manipulation commands available in FFX.

Command	**Description**
Cart2Frac	Convert from Cartesian to fractional coordinates
Frac2Cart	Convert from fractional to Cartesian coordinates
SaveAsP1	Expands a crystal to P1 or a replicated unit cell
SaveAsXYZ	Save the system as an XYZ file
SaveAsQE	Create a Quantum ESPRESSO input script
ImportCIF	Import a system from a CIF file
MoveIntoUnitCell	Move all molecules into the unit cell
xray.MTZInfo	Log information for an MTZ file
xray.CIFtoMTZ	Convert diffraction data from CIF format to MTZ

Software Organization. FFX is composed of nine Java packages organized by functional themes and capabilities. The nine packages are Parallel Java (PJ), Utilities, Numerics, Crystal, OpenMM, Potential, Algorithms, Refinement, and User Interfaces (UI). The organization of packages and their dependencies within FFX are shown in Figure 2.

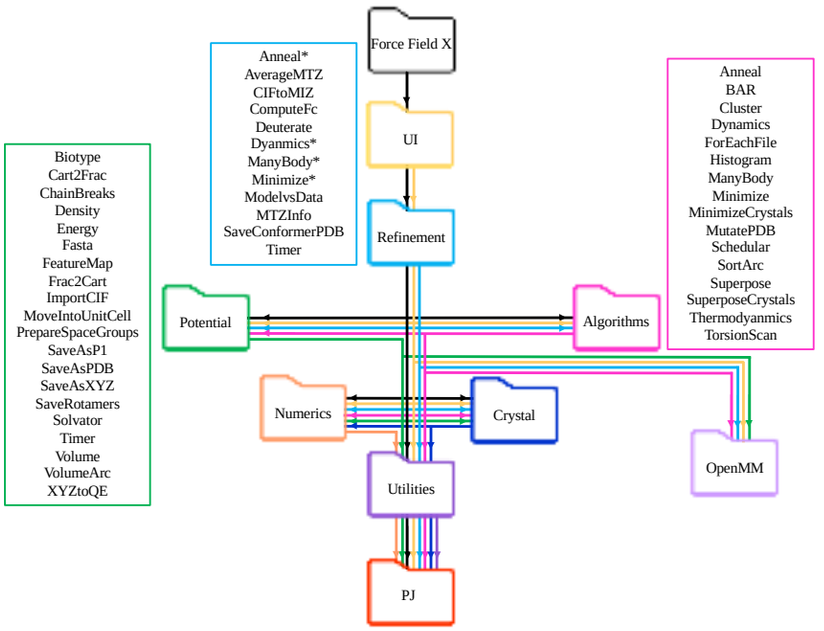

Figure 2. A flowchart representation of the package dependencies in FFX. Each package is indicated by a separate color and the lines between packages indicate a dependency relationship. For example, most colored lines connect to the PJ package as all packages depend on PJ except for OpenMM. Commands available in the Potential, Algorithms, and Refinement packages are displayed in the corresponding color-coded box. For Refinement commands, those with an * indicates both X-ray and Real Space versions of the command are available, while those without an * currently have only an X-ray version.

Parallel Java[60,61] provides APIs for both shared memory (SM) parallelization using threads and a message-passing interface (MPI) for parallelization across JVM processes that are executing on the same node and/or between nodes. The former SM parallelization leverages concepts analogous to those defined by OpenMP, including support for parallelization of “for” loops, atomic operations, and scheduling of tasks. The PJ MPI approach defines its own scheduling and communication (*e.g.*, “mpirun” is unnecessary) without any dependencies beyond the Java Runtime Environment (JRE). The Parallel Java package is used to accelerate operations in most other packages (*e.g.*, 3D FFTs, force field energy evaluations, and replica-based sampling strategies). Therefore, all packages have PJ as a dependency. Our fork of the original PJ code by Alan Kaminsky contains several changes geared toward freeing SM hardware threads that are no longer in use and enhanced logging of MPI communications[60,61].

The Utilities package provides simple functionality for file handling (*e.g.*, file copying and file naming conventions) and the implementation of system properties. Properties in FFX are built on the JVM standards for system properties. These properties (key-value pairs) are used to specify calculation details that vary between simulation goals. All packages except for PJ and OpenMM depend on the Utilities package.

The Numerics package implements common mathematical methods and numerical recipes required for calculations available within FFX. The Numerics package computes real and complex 1D[62,63] and 3D fast Fourier transforms (FFT), offers both float (single precision) and double (double precision) vector math libraries, and a range of special functions such as Erf/Erfc and Modified Bessel functions. Numerics offers a novel family of multipole tensor recursion[64] algorithms for Coulomb, Ewald, Thole, and GK interactions using both Cartesian and quasi-internal[65] coordinate frames. Finally, the Numerics package implements atomic operations on arrays, a limited-memory BFGS[66-69] optimizer, and support for uniform b-splines[70]. The Crystal, Potential, Algorithms, Refinement, and UI packages each depend on the Numerics package.

The Crystal package gives FFX the ability to perform operations on all 230 crystallographic space groups during evaluation of a force field potential energy and during crystallographic refinement against X-ray and/or neutron diffraction data. It provides methods for application of the minimum image convention, symmetry operators, conversion between cartesian and fractional coordinates, and the storage of reflection lists from diffraction experiments[71]. The Potential, Algorithms, Experiment, and UI packages each depend on the Crystal package.

The Potential package provides support for evaluating the potential energy of atomic resolution molecular systems using fixed partial atomic charge force fields, polarizable atomic multipole force fields, and preliminary support for using neural network potentials that are based on PyTorch[72]. When a molecular system is loaded from a file, an instance of the MolecularAssembly class is instantiated together with creation of an associated

ForceFieldEnergy. The latter provides methods to compute the potential energy of the system and optionally, its gradient for use with simulation methods defined in the Algorithms package described below. The Potential package supports calculation of long-range electrostatics using particle-mesh Ewald[36-38] (PME) summation with novel support for space group symmetry[1] and offers continuum treatment of solvent via our GK model[3,33-35]. The FFX commands defined within the Potential package are listed in Figure 1. Some notable commands include *Energy* to calculate the potential energy of a system, *Solvator* to solvate a system into a water box with or without free salt, and the *SaveAs* family of commands to convert systems between file types. The Algorithms, Refinement, and UI packages depend on the Potential package.

The Algorithms package contains optimization and sampling methods that operate on the potential energy functions defined in the Potential package. The Algorithms package commands are listed in Figure 2. A family of local optimization commands (*Minimize*, *CrystalMinimize,* and *PhMinimize*) leverage the L-BFGS method defined in the Numerics package. The Algorithms package also offers global optimization methods based on simulated annealing (*Anneal*) and via our many-body versions[4,11] of dead-end elimination[41] and Goldstein elimination[73] (*ManyBody*). The *Dynamics* command executes molecular dynamics via integrators that include velocity Verlet, Beeman[74], stochastic dynamics[75,76], and reversible reference system propagation algorithm (r-RESPA)[77,78]. In the absence of stochastic dynamics, temperature control is available via thermostats by Berendsen[79] and Bussi[80]. Constant pressure is achieved using a novel Monte Carlo barostat[81] that respects space group constraints. The *Thermodynamics* command computes free energy differences via an alchemical path defined by a state variable (λ) either by sampling an array of windows at fixed λ values (followed by application of the Bennett Acceptance Ratio method[82]) or using a unique implementation of the orthogonal space tempering[39,83] method that supports both polarizable force fields and space group symmetry[6-8,84]. The Refinement and User Interface packages depend on the Algorithms package.

The Refinement package implements target functions that combine a potential energy function defined in the Potential package (*e.g.*, a fixed charge or polarizable force field) with a function that compares a molecular model (*e.g.*, its atomic coordinates, b-factors and/or occupancies) to either diffraction data in reciprocal space or a scalar field (*e.g.*, electron density) in real space[10]. This includes a novel bulk scattering model that is differentiable with respect to atomic coordinates[9] and the unique ability to evaluate long-range electrostatics using the rigorous PME method[1]. Once the overall target function is defined, then most of the optimization and sampling methods of the Algorithms package can be employed for model refinement (*e.g.*, local minimization, molecular dynamics, many-body side-chain optimization, and simulated annealing). Refinement against X-ray diffraction data, neutron diffraction data, or joint X-ray/neutron diffraction data sets is supported. Global optimization of side-chain conformations against either a reciprocal space or real space target function is supported[11]. The available commands in the Refinement package are listed in Figure 2. Those with an asterisk (*) have both reciprocal and real space versions. The User Interface package depends on the Refinement package.

The User Interface package implements both command line and graphical user interfaces. This package depends on all other packages. We will elaborate on FFX functionalities in more explicit detail later in this article.

External Libraries

FFX leverages a variety of external libraries to perform functions like CIF file parsing, parallelization, and bioinformatics methods. These packages include CIF tools[85], BioJava[86], the Chemistry Development Kit (CDK)[87], TornadoVM[88], Parallel Java[60], the picocli (a command line interface package)[89], the Groovy language, and the GraalVM Python implementation. CIF tools give FFX the ability to read and write CIF files from the Cambridge Structural Database[59]. The BioJava library supports local installations of the PDB, can load protein and nucleic acid sequences stored in FASTA format, and offers algorithms for structural alignment. The Chemistry

Development Kit (CDK) is a collection of modular libraries for processing chemical information, including support for parsing files in SMILES, SDF, and Mol2 format. CDK also allows substructure and SMARTS pattern matching, and fingerprint methods for similarity searching. TornadoVM extends the JRE by offering programming constructs to facilitate translating a subset of Java code into PTX (CUDA), OpenCL, or SPIRV (Intel Level Zero) backends. As mentioned previously, the Parallel Java library implements both "OpenMP Style" shared memory coding constructs and a platform independent implementation of the canonical message passing interface operations for parallelization across processes. Picocli is an annotation driven library for creating command line applications along with documentation in HTML, PDF, or Unix man page formats. The Groovy scripting language is used to quickly prototype new ideas and create FFX commands. Finally, with the emergence of a Python 3.10 implementation that runs on the JVM from the GraalVM team, FFX now also fully supports execution of Python scripts in a cross-platform manner.

Potential Energy Functions

FFX supports a variety of potential energy functions that includes both fixed partial charge and polarizable atomic multipole force fields. There is support for implicit solvents consisting of cavitation, dispersion and generalized Kirkwood contributions, energy terms for refinement against X-ray and/or neutron diffraction data, and energy terms for refinement against real space maps (*e.g.*, from either CryoEM or diffraction experiments). FFX also offers unique support for a family of dual-topology potential energy functions that facilitate computing free energy differences due to chemical modifications in the AMOEBA force field (*e.g.*, in the context of relative binding affinity or relative hydration free energy), between two crystal polymorphs, or between alternative force field models. Force field parameter files adopt Tinker conventions when possible, however, support for experimental refinement and dual-topology algorithms are unique to FFX.

Force Field Models. The overall force field functional form combines bonded and non-bonded interactions.

$$U_{\text{Force Field}} = U_{\text{Bonded}} + U_{\text{Non-Bonded}}$$

Equation 1

The bonded energy may include bond stretching, Urey-Bradley stretching, angle bending, bond-angle cross-term, out-of-plane bending, improper torsion, torsional angle, stretch-torsion cross-term, angle-torsion cross-term, and/or a torsion-torsion cross-term.

$$U_{\text{Bonded}} = U_{\text{Bonds}} + U_{\text{Urey-Bradley}} + U_{\text{Angles}} + U_{\text{Bond-Angles}} + U_{\text{Out-of-Plane Bends}} + U_{\text{Torsions}} + U_{\text{Improper-Torsions}} + U_{\text{Stretch-Torsions}} + U_{\text{Angle-Torsions}} + U_{\text{Torsion-Torsions}}$$

Equation 2

However, no currently supported force field includes all ten bonded terms. For example, the OPLS-AA[90] and OPLS-AA/L[91] force fields include the following four terms: bond stretching, angle bending, improper torsion, and torsional angle. On the other hand, a simulation that utilizes the AMOEBA nucleic acid force field[16] in explicit water[12] includes all terms except improper torsions. The non-bonded terms may include van der Waals interactions (using either 6-12 Lennard-Jones or Buffered-14-7 functional forms), permanent electrostatics based on fixed atomic charges, or multipoles (truncated at quadrupole order) and polarization energy via induced dipoles.

$$U_{\text{Non-Bonded}} = U_{\text{vdW}} + U_{\text{Elec}}^{\text{Permanent}} + U_{\text{Elec}}^{\text{Induced}}$$

Equation 3

Currently supported force fields, in addition to those mentioned above, include Amber94[92], Amber96[93], Amber99[94], Amber99sb[95], Charmm22[96,97], Charmm22 with CMAP correction[98], and the AMOEBA model for small organic molecules[14] and proteins[15].

Implicit Solvent

FFX implements implicit solvents for applications where the use of explicit water molecules is cumbersome, including the repacking of protein side chains and the refinement of biomolecular models against experimental data. The implicit solvent is generally formulated as a sum of three

free energy differences defined by a thermodynamic cycle: cavitation, dispersion, and electrostatics contributions. The cycle describes the transfer of the biomolecular system between vacuum into solvent phases. The combination of cavitation and dispersion free energy differences is usually referred to as the non-polar contribution[59-62], while the latter electrostatic contribution is based on an analytic approximation to the Poisson equation as described below.

The Generalized Born (GB) model approximates the polar, electrostatic term for fixed partial charge force fields[99,100]. GB is formulated as a sum over pairwise and self-interactions to approximate the electrostatic solvation free energy difference (ΔG_{GB}).

$$\Delta G_{GB} = \frac{1}{2}\left(\frac{1}{\varepsilon_s} - \frac{1}{\varepsilon_h}\right)\sum_{i,j}\frac{q_i q_j}{f_{ij}}$$

Equation 4

where ε_s is the permittivity of the solvent, ε_h is the permittivity of the homogenous reference state, and q_i and q_j are partial charges. The original form of the generalizing function, f_{ij} is defined in Equation 5.

$$f_{ij} = \sqrt{r_{ij}^2 + a_i a_j \exp\left(-r_{ij}^2/c a_i a_j\right)}$$

Equation 5

where r_{ij} is the separation distance between atoms, a_i and a_j are effective Born radii[101], and c is a constant to control the transition from the Born regime to the Coulomb's law regime. To support the polarizable atomic multipole AMOEBA force field, GB concepts were extended to define the GK model that handles multipole moments of arbitrary degree[33]. The GK monopole term ($\Delta G_{GK}^{(q,q)}$) is equivalent to GB (Equation *4*). To calculate the interactions between all permanent dipole moments, the GK dipole term is defined as

$$\Delta G_{GK}^{(\mu,\mu)} = \frac{1}{2}\left[\frac{1}{\varepsilon_h}\frac{2(\varepsilon_h - \varepsilon_s)}{2\varepsilon_s + \varepsilon_h}\right]\sum_{i,j}\mu_{i,\alpha}\mu_{j,\beta}\left[\frac{3r_\alpha r_\beta g_{ij}}{f_{ij}^5} + \frac{\delta_{\alpha\beta}}{f_{ij}^3}\right]$$

Equation 6

where $\boldsymbol{\mu_i}$ and $\boldsymbol{\mu_j}$ are permanent atomic dipole moments, the subscripts α and β denote the use of the Einstein summation convention, $\delta_{\alpha\beta}$ is the Kronecker delta, the separation along the α dimension is given by $r_\alpha = r_{j,\alpha} - r_{i,\alpha}$, and the chain rule term (g_{ij}).

$$g_{ij} = \frac{\exp\left(-r_{ij}^2/ca_ia_j\right)}{c} - 1$$

Equation 7

Both the GB and GK equations use effective radii, rather than the intrinsic radius of each atom, which represent the degree of burial within the solute. An atom that is deeply buried within the center of a protein has a larger effective radius than a surface exposed atom of the same type. An isolated atom's (*i.e.,* an ion) effective radius will approach its intrinsic atomic radius.

FFX calculates effective radii by combining the analytic Hawkins, Cramer, and Truhlar (HCT) pairwise descreening approximation[102] with the solvent field approximation (SFA) proposed by Grycuk[103]. Contributions to effective radii due to interstitial spaces (spaces within a system where a water molecule cannot fit) are also included due to their importance when modeling proteins. The interstitial space corrections include a pairwise "neck" between nearby atoms[104] and a hyperbolic tangent (tanh) function[105] to help smoothly scale up the effective radius of an atom as it becomes more deeply buried. Currently, the FFX implementation of GK implicit solvent has been validated for protein simulations with work ongoing to support nucleic acids[3].

Dual-Topology Framework

Interpolation Between Force Fields. FFX implements two indirect free energy (IFE), or bookending methods[7,8]. These methods seek to correct thermodynamic quantities obtained with a relatively low-resolution potential energy function (***e.g.***, a molecular mechanics force field) to be consistent with a higher-resolution potential energy function (***e.g.***, a polarizable molecular mechanics force field, neural network, or QM/MM potential)[106]. The first approach is called the dual force field (DFF) method[7]. DFF creates an alchemical path between a relatively expensive potential at one end (*i.e.*, the polarizable AMOEBA model U_{AMOEBA}) and a relatively inexpensive

potential at the other end (*i.e.*, a fixed partial charge force field U_{FC}). The potential energy along the path is then given by

$$U_{DFF} = \lambda * U_{\mathrm{AMOEBA}}(\mathbf{x}) - (1-\lambda)U_{\mathrm{FC}}(\mathbf{x})$$

Equation 8

where the λ is a state variable that ranges between 0 and 1 to parameterize the transition between force field resolutions and $\mathbf{x}$ are the atomic coordinates. This approach has been successfully used to compute the free energy difference for the sublimation of small organic molecules[7] and more recently, in the context of computing relative anhydrous–hydrate stability[44].

Despite these successes, a limitation of the DFF approach is that both end states must have the same degrees of freedom and constraints (*e.g.,* bonds and angles must either be flexible or rigid under both potentials). DFF also has difficulty converging the free energy difference for large systems. This is due to the size extensive nature of the calculation in tandem with relatively large contributions from slight differences in the bonded terms' equilibrium values. The latter limitation is partially addressed by the second IFE method in FFX called Simultaneous Bookending (SB)[8]. SB couples two DFF simulations running in opposite directions – the first coarsens the resolution of the system and the second refines the resolution back to the more accurate potential energy function -- to compute the free energy correction for the entire transformation in one step rather than using two IFE simulations (*i.e.*, one at each end of the thermodynamic path). SB does not entirely solve convergence issues, but it does allow for an approximation based on distance to the site of the original alchemical transformation. Atoms distal to the site of an alchemical transformation can be pinned or constrained to share identical coordinates. This constraint dramatically improves sample convergence because many energy terms do not contribute to the partial derivative of the total potential energy with respect to λ and thus do not contribute to the free energy difference. The SB potential energy along the alchemical path is given by a weighted sum of four force field energy contributions (or two DFF potentials).

$$U_{SB}(\acute{\mathbf{x}}_C, \acute{\mathbf{x}}_R, \mathbf{x}_S) = \lambda * \left[U_{\mathrm{A,FC}}(\acute{\mathbf{x}}_C, \mathbf{x}_S) + U_{\mathrm{B,AMOEBA}}(\acute{\mathbf{x}}_R, \mathbf{x}_S)\right] + (1-\lambda) * \left[U_{\mathrm{B,FC}}(\acute{\mathbf{x}}_R, \mathbf{x}_S) + U_{\mathrm{A,AMOEBA}}(\acute{\mathbf{x}}_C, \mathbf{x}_S)\right]$$

Equation 9

where $\acute{\mathbf{x}}_S$ are the shared (or pinned) coordinates common to both sides of the SB simulation (*e.g.*, atoms distal from the site of alchemical transformation) while $\acute{\mathbf{x}}_C$ and $\acute{\mathbf{x}}_R$ are independent degrees of freedom near the alchemical transformation for the "coarsen" and "refine" steps, respectively.

Overly aggressive coordinate constraints cause the calculated free energy correction to diverge from the actual free energy correction due to the approximation that both systems share an identical phase space. On the other hand, overly conservative constraints render the correction increasingly expensive to converge. The SB approach was used to correct the binding free energy differences for a set of divalent cation binding proteins to within statistical uncertainty of the true calculated AMOEBA values[8]. Compared to prior IFE methods, the SB approach allowed an order of magnitude more atoms to be converted between resolutions. Future IFE efforts would benefit from force field resolutions with identical bonded terms (*e.g.*, between the fixed partial charge and polarizable atomic multipole resolutions) such that only non-bonded terms contribute to the corrections.

Interpolation Between Polymorphs. A recent addition to the dual topology framework, described here for the first time, employs a symmetry operator to interpolate between polymorphs and thereby estimate their free energy difference. The dual polymorph potential energy is defined as:

$$U_{DP}(\mathbf{x}_A, \lambda) = (1-\lambda) * U_A(\mathbf{x}_A, \lambda) + \lambda * U_B(S_{A\to B}(\mathbf{x}_A), \lambda)$$

Equation 10

where the coordinates of the first polymorph ($\mathbf{x_A}$) are used to generate those of the second using a symmetry operator ($S_{A\to B}$) defined using the Progressive Alignment of Crystals algorithm

presented in the following section on Structure Manipulation. At $\lambda = 0$, the molecule(s) within the asymmetric unit experience the crystalline environment defined by the space group of polymorph A. At $\lambda = 1$, they experience the crystalline environment defined by the space group of polymorph B. At intermediate values of the state parameter (λ), the symmetry mate interactions are smoothly transformed as depicted in Figure 3. At each molecular dynamics step, the inverted symmetry operator is applied to rotate forces produced by the second polymorph (B) back into the frame of the first polymorph (A). Overall, this relative free energy difference approach for polymorphs offers performance advantages relative to taking the difference between two absolute free energy differences (*i.e.*, analogous to the advantages of relative binding free energy differences).

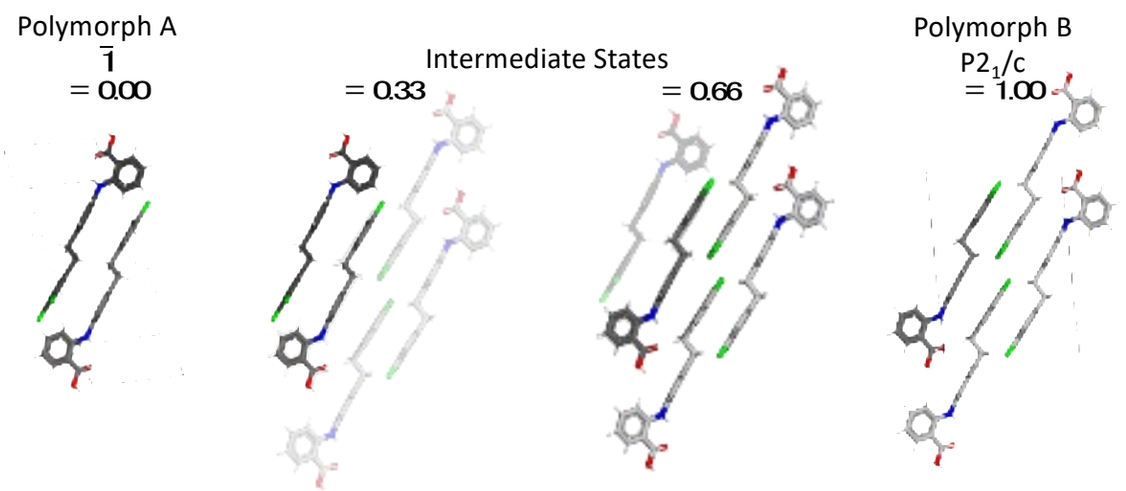

Figure 3. Depiction of an alchemical path connecting two crystal polymorphs defined by different space groups via a dual topology framework (Equation 10). Symmetry mates produced via the $P\bar{1}$ symmetry operators of polymorph A are shown with carbon atoms colored black and those produced by the $P2_1/c$ symmetry operators of polymorph B are light gray. The atoms of the asymmetric unit of polymorph A are mapped via a custom symmetry operator into the asymmetric unit of polymorph B (asymmetric unit carbon atoms are colored gray for each state).

Structure Manipulation

FFX contains methods to alter systems as needed. *MutatePDB* changes the identity of a chosen amino acid to an alternative identity. *Solvator* creates a periodic box of water around the input system and optionally adds explicit counterions. The *Superpose* command calculates the

coordinate root mean square deviation (RMSD) between two systems via quaternion superposition. A unique *SuperposeCrystals* command quantifies the packing similarity between two crystals. This is performed using the Progressive Alignment of Crystals (PAC) algorithm, which calculates an atomistic RMSD for the aligned molecular subclusters of each crystal[57]. The input options the user can select include the number of asymmetric units in each subcluster, the selection criteria for molecules in the subcluster, and atom(s) to be excluded from the comparison. *SuperposeCrystals* can save crystal systems that are within a user specified RMSD from a desired crystal or to write out a matrix of RMSD values for clustering and/or filtering out similar crystals within an ensemble. Furthermore, PAC can accumulate the rotations and translations into an overall symmetry operator to map moleculars between crystal polymorphs as discussed in the prior section on "Interpolation Between Polymorphs".

Local Optimization

FFX currently has two methods for local optimization: the steepest descent method and the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method. The latter is a quasi-Newton approach that minimizes a nonlinear function by approximating the inverse Hessian matrix[107,108]. Both methods are available via the *Minimize* command to optimize atomic coordinates for a given potential energy function. The *MinimizeCrystals* command additionally supports optimization of lattice parameters, while the *MinimizePh* command supports local optimization of titration and tautomer states. Finally, the *xray.Minimize* and *realSpace.Minimize* commands facilitate optimization of coordinates, b-factors and/or occupancies against experimental diffraction or real space data, respectively.

Global Optimization

Simulated Annealing. FFX implements simulated annealing functionality via the *Anneal* command[52,109]. The command allows selection of the heating and cooling schedule, the molecular

dynamics protocol and simulation length at each temperature, and selection of the atomic degrees of freedom that should remain fixed. Although the simulated annealing approach can locate the global minimum of a target function, its success depends on the simulations at each temperature being of sufficient length. Versions of the method are available for refinement against both experimental diffraction data (*xray.Anneal*) and real space maps (*realSpace.Anneal*)[110].

Methods Based on a Many-Body Expansion. Under a many-body potential, such as a polarizable force field[15], implicit solvent[3], and/or X-ray diffraction target[11], the total energy of the system $E(\mathrm{r})$ can be defined to arbitrary precision using a many-body expansion.

$$E(\mathrm{r}) = E_{\mathrm{env}} + \sum_i E_{\mathrm{self}}(r_i) + \sum_i \sum_{j>i} E_2(r_i, r_j) + \sum_i \sum_{j>i} \sum_{k>j} E_3(r_i, r_j, r_k) + \cdots$$

Equation 11

where E_{env} is the energy of the environment (*e.g.*, a protein backbone and any residues that are not being optimized). $E_{\mathrm{self}}(r_i)$ is the self-energy of residue i that includes its intra-molecular bonded energy terms and non-bonded interactions with the backbone. $E_2(r_i, r_j)$ is the 2-body non-bonded interaction energy between residues *i* and *j*, and $E_3(r_i, r_j, r_k)$ is the 3-body non-bonded interaction energy between residues *i*, *j*, and *k*. The self, two-body, and three-body energy terms from Equation 11 are calculated as follows, where $E_{\mathrm{env/sc}}$ is the total energy of the environment and the side chain(s) of the selected residue(s).

$$E_{self} = E_{\mathrm{env/sc}}(r_i) - E_{env}$$

Equation 12

$$E_2(r_i, r_j) = E_{\mathrm{env/sc}}(r_i, r_j) - E_{\mathrm{self}}(r_i) - E_{\mathrm{self}}(r_j) - E_{\mathrm{env}}$$

Equation 13

$$E_3(r_i, r_j, r_k) = E_{\text{env/sc}}(r_i, r_j, r_k) - E_{\text{self}}(r_i) - E_{\text{self}}(r_j) - E_{\text{self}}(r_k) - E_2(r_i, r_j) - E_2(r_i, r_k) - E_2(r_j, r_k) - E_{\text{env}}$$

Equation 14

In the past, global optimization algorithms over a discrete permutation space (*i.e.*, side chain conformations and/or amino acid identity) using a many-body expansion were limited to truncation at pairwise interactions[41,73]. We demonstrated that the pairwise dead-end elimination and Goldstein elimination criteria can be modified to include 3-body (or higher) energy terms[11]. However, computing the self, 2-body, and 3-body energy terms as a function of rotamer conformation is computationally expensive. To address this challenge, we demonstrated two complementary parallelization approaches in FFX, including 1) use of MPI parallelization to distribute terms among multiple processes, and 2) use of the OpenMM API to perform energy evaluations on NVIDIA GPUs *via* CUDA kernels[4].

In addition to extending the Goldstein elimination criteria and implementing parallelization strategies, four approximations were introduced to improve computational performance. In the context of rotamer optimization, the expansion can often be truncated at pairwise terms due to damping of 3-body and higher order terms by the generalized Kirkwood implicit solvent. However, previous work also demonstrated that inclusion of 3-body terms is sometimes necessary *(i.e.,* in the absence of implicit solvent or when using an X-ray diffraction target function)[11]. The second approximation employs a distance cutoff to exclude interactions where the closest rotamers for a residue pair are more than 3 Å apart or for residue triples that are more than 2 Å apart. Pruning removed rotamers with self-energies 25 kcal/mol or more above the lowest self-energy of a residue, prior to calculation of 2-body energies. The pruning criterion is based on the heuristic observation that rotamers with such an unfavorable self-energy *(e.g.*, due to an atomic clash with backbone atoms) are not found in well-packed structures. The final approximation imposed a 3D grid over the protein, followed by optimization within each subdomain (cube) of the grid, rather

than including all protein residues simultaneously. The repacking algorithm is a provable global optimizer within a single subdomain of the grid, but not for the whole protein because coordinated changes between subdomains are neglected.

Periodic Systems, Space Group Symmetry, and Neighbor Lists

The conditional convergence of Coulomb's law under periodic boundary conditions can be treated by splitting the electrostatic potential into a strictly convergent short-ranged real space contribution and a smooth periodic contribution. The smooth periodic contribution is strictly convergent in Fourier space as described by Ewald[111]. For crystals of small organic molecules, both space group symmetry and handling of the minimum image convention require special consideration when building the neighbor list as shown in **Error! Reference source not found.** for Compound 23. In FFX, both issues are handled in a uniform fashion based on permuting space group symmetry operators with the translational operators needed to generate replicated copies of the unit cell (*i.e.*, to generate an overall cell that is larger than twice the cutoff used during building of the neighbor list). In direct space, the non-bonded pairwise loops over neighbors (*i.e.*, for van der Waals or electrostatics interactions) leverage an additional outer loop over the permuted symmetry operators. FFX then utilizes a customized version of the smooth particle mesh Ewald (PME) algorithm that handles both space group symmetry and small unit cells[112]. For NPT simulations, the number of replicates cells is adjusted dynamically to accommodate shrinking or growing unit cells.

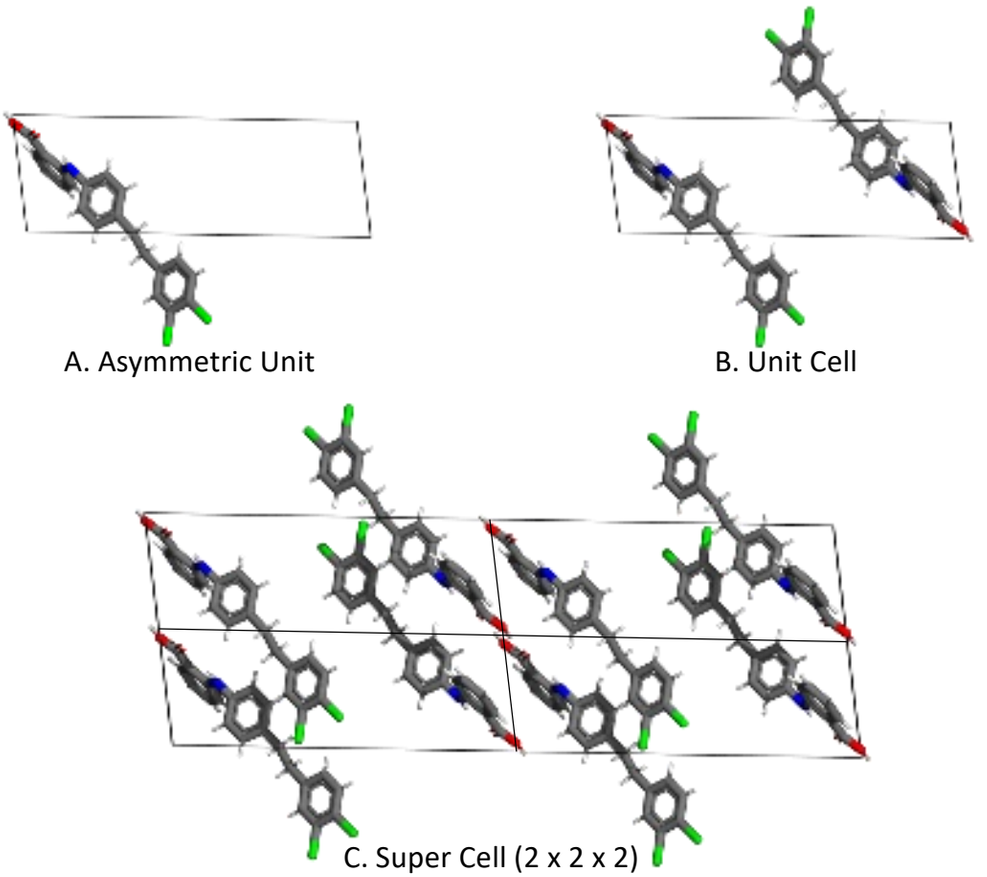

Figure 4. FFX supports all 230 space groups, the use of unit cells whose interfacial radius is less than half of the nonbonded cutoff distance, and the combination of small unit cells with space group symmetry.

Particle Mesh Ewald Summation

As a part of FFX's potential energy capabilities, a general implementation of smooth particle-mesh Ewald summation (PME) is included to avoid using electrostatic cut-offs and boost performance[36,37]. The FFX implementation of PME for multipoles builds on the work of Sagui *et al.*[38] by adding unified support for symmetry operators for all 230 crystallographic space groups, replicates operators for small unit cells (*i.e.*, interactions with neighboring unit cells are required to satisfy the real space cut-off), and their combination[1]. Here, we will emphasize notable changes

in the algorithm compared to that available in other simulation packages due to inclusion of symmetry operators for both the direct and reciprocal space terms.

Real Space Energy. The real space permanent atomic multipole interaction energy is tuned by the Ewald parameter (β).

$$U_{\text{real}} = \frac{1}{2}\sum_{s_j=1}^{n_{r,}*}\sum_{i=1}^{n_a}\sum_{j=1}^{n_a} L_i(\mathbf{I})L_j\left(\mathbf{R}_{s_j}\right)\frac{\operatorname{erfc}\left(\beta\left|\mathbf{r}_i-\left(\mathbf{R}_{s_j}\mathbf{r}_j+\mathbf{t}_{s_j}\right)\right|\right)}{\left|\mathbf{r}_i-\left(\mathbf{R}_{s_j}\mathbf{r}_j+\mathbf{t}_{s_j}\right)\right|}$$

Equation 15

where the outer summation is over n_r space group symmetry and/or replicates operators, and the inner summations are over n_a atoms in the asymmetric unit. The distance from an atom in the asymmetric unit at $\mathbf{r}_i$ to another atom at $\mathbf{R}_{s_j}\mathbf{r}_j+\mathbf{t}_{s_j}$ is a function of operator s_j with Cartesian rotation matrix $\mathbf{R}_{s_j}$ and translation vector $\mathbf{t}_{s_j}$.The asterisk on the outer summation indicates that $i=j$ interactions are neglected and masked interactions are respected for the identity symmetry operator ($s_j=1$ by convention). Finally, the operators L_i and L_j are

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

Equation 16

and

$$L_j(\mathbf{R}) = q_j - \left(\mathbf{R}\mathbf{d}_j\right)_\alpha\nabla_{i,\alpha} + \left(\mathbf{R}\boldsymbol{\Theta}_j\mathbf{R}^{\mathrm{t}}\right)_{\alpha\beta}\nabla_{i,\alpha}\nabla_{i,\beta}\frac{1}{3}$$

Equation 17

where charge (q), dipole ($\mathbf{d}$), and traceless quadrupole ($\boldsymbol{\Theta}$) moments are operated on by a Cartesian rotation matrix. $\nabla_{i,\alpha}$ is one component of the del operator acting at $\mathbf{r}_i$, $\alpha \in \{x,y,z\}$, and $\{\alpha,\beta\}$ represent use of the Einstein summation convention for summing over tensor elements. A replicates super cell with $l\times m\times n$ copies of the unit cell and n_s space group symmetry operators

will require $n_r = n_s \times l \times m \times n$ total symmetry operators. For large crystals, replicated copies of the unit cell are not required to satisfy the real space cut-off and $n_r = n_s$.

Reciprocal Space Energy. The PME reciprocal space multipolar electrostatic energy U_{rec} for a unit cell[38] is given by

$$U_{\text{rec}}^{\text{U.C.}} = \frac{1}{2} \sum_{m_1=0}^{K_1-1} \sum_{m_2=0}^{K_2-1} \sum_{m_3=0}^{K_3-1} Q^R(\mathbf{m}) \cdot (G^R * Q^R)(\mathbf{m})$$

Equation 18

where Q^R is the reciprocal lattice grid of dimension $\{K_1, K_2, K_3\}$ populated with multipoles using B-splines (θ_p).

$$\begin{aligned} Q^R(k_1, k_2, k_3) = \sum_{s_j=1}^{n_s} \sum_{i=1}^{n_a} \sum_{n_1,n_2,n_3} L_i\left(\mathbf{R}_{s_j}\right) [\theta_p(K_1 u_{1i} - k_1 - n_1 K_1) \\ \times \theta_p(K_2 u_{2i} - k_2 - n_2 K_2) \\ \times \theta_p(K_3 u_{3i} - k_3 - n_3 K_3)] \end{aligned}$$

Equation 19

where L_i was given in Equation 16, $\mathbf{u}_i$ are the fractional coordinates of atom *i*, the summation over all integers $\{n_1, n_2, n_3\}$ is finite due to the local support of B-splines, and G^R is the discrete Fourier transform of the coefficients arising from the structure factor. The reciprocal space electrostatic energy for the asymmetric unit is

$$U_{\text{rec}}^{\text{A.U.}} = \frac{1}{2} \sum_{m_1=0}^{K_1-1} \sum_{m_2=0}^{K_2-1} \sum_{m_3=0}^{K_3-1} Q_{\text{A.U.}}^R(\mathbf{m}) \cdot (G^R * Q^R)(\mathbf{m})$$

Equation 20

where the multipoles experiencing the reciprocal space potential are limited to the asymmetric atoms.

$$Q^{R}_{\text{A.U.}}(k_1,k_2,k_3) = \sum_{i=1}^{n_a} \sum_{n_1,n_2,n_3} L_i\,(\mathbf{I})\,[\theta_p(K_1 u_{1i} - k_1 - n_1 K_1) \times \theta_p(K_2 u_{2i} - k_2 - n_2 K_2) \times \theta_p(K_3 u_{3i} - k_3 - n_3 K_3)]$$

Equation 21

Both the grid dimensions and spline order can be set by via FFX properties. Application of the multipoles onto the FFT grid is parallelized spatially using 3D domains[1], over 2D slices of the grid, or over 1D rows. The convolution is then performed by a parallelized 3D FFT[62,63], followed by a pointwise multiplication in reciprocal space, and finally the inverse 3D FFT.

Polarization Algorithms

Definition of the Self-Consistent Field. To calculate the energy and gradient for a polarizable force field (*e.g.*, AMOEBA), we must solve for the self-consistent field (SCF) that induces dipoles. FFX supports multiple iterative SCF solvers. The mutual (or self-consistent) induced dipoles ($\mathbf{u}$) are a function of the isotropic atomic polarizability of each atom ($\boldsymbol{\alpha}$) and the total electric field ($\mathbf{E}$) at each site.

$$\mathbf{u} = \boldsymbol{\alpha}\mathbf{E}$$

Equation 22

where $\mathbf{u}$ and $\mathbf{E}$ are both vectors of dimension 3*n* and the polarizability $\boldsymbol{\alpha}$ is a diagonal 3*n* *x* 3*n* matrix. The total electric field can be broken into a "direct" contribution from permanent multipoles ($\mathbf{E}_{\text{direct}}$) and the contribution from induced dipoles.

$$\mathbf{u} = \boldsymbol{\alpha}(\mathbf{E}_{\text{direct}} + \mathbf{T}\mathbf{u})$$

Equation 23

where $\mathbf{T}$ is a *3n* x *3n* matrix of interaction tensor elements that operate on the induced dipole vector to produce the induced field at atomic sites. The induced dipoles can be solved analytically.

$$\mathbf{u} = \mathbf{C}^{-1}\mathbf{E}_{\text{direct}}$$

Equation 24

where explicit inversion of the matrix $\mathbf{C} = \left(\boldsymbol{\alpha}^{-1} - \mathbf{T}\right)$ scales $O(n^3)$. Better scaling is achieved through iterative methods including successive over-relaxation (SOR), a preconditioned conjugate gradient (PCG) solver, or an approach based on optimized perturbation theory (OPT). Each method will be briefly discussed here, while more details are available from Lipparini *et al.*[113]

Successive Over Relaxation (SOR). The SOR technique is a variant of the Gauss-Seidel method for solving a linear system of equations. It incorporates the previous solution of the system of linear equations with a variable called the relaxation factor (0.7 by default). Although the convergence rate can be improved slightly by tuning the relaxation factor to the system of interest, SOR requires significantly more iterations than the Preconditioned Conjugate Gradient method described below to achieve the same convergence criteria. By default, the SCF convergence criteria is to reduce the change in induced dipole magnitude between iterations below a threshold of 1.0e^{-6} RMS Debye.

Preconditioned Conjugate Gradient (PCG). The objective of the PCG technique is to minimize the residual found by moving all components of the linear system $\mathbf{Ax} = \mathbf{b}$ to the right-hand side (r.h.s.). The notation for AMOEBA is achieved by rearranging Equation 23 to give the residual as $\mathbf{r} = \mathbf{E}_{\text{direct}} - \mathbf{Cu}$. A preconditioner is used to improve the condition number of the matrix $\mathbf{C}$. In FFX, the preconditioner is based on using a short real space cut-off with a default value of 4.5 Å (with no reciprocal space contribution to the field). The PCG method typically reduces the RMS Debye change between iterations by an order of magnitude each cycle (*i.e.*, just six or seven cycles are needed to reach the default convergence criteria).

FFX allows users to define the polarization model, where the default value of "mutual" (*i.e.*, SOR or PCG is used to iteratively converge the SCF) can be changed to "direct" or "none". The

"direct" option includes the field due to permanent multipoles, but not the field due to induced dipoles themselves (*i.e.*, the 2nd term on the r.h.s. of Equation 23 is not included). Selecting "none" eliminates the polarization energy term from the potential energy and is useful for relaxing poor starting coordinates whose SCF is unstable.

Dynamics Methods

Integrators and Controls. Several integrators are implemented to propagate degrees of freedom based on Newton's equations of motion during molecular dynamics (MD) simulations. Velocity Verlet provides the simplest integration scheme that is numerically stable and time-reversible[114,115]. The closely related Beeman[74] algorithm for integration produces an identical trajectory to velocity Verlet with a modification that calculates velocities more accurately and better conserves energy. FFX implements a reversible-RESPA integrator[116] that offers multiple-time step simulations through the separation of long-range force calculations from the short-range. The short-range forces are calculated at each time step by a position Verlet algorithm. The position Verlet offers greater stability than velocity Verlet if large time steps are used. The long-range forces are calculated at *n* time steps, reducing the computational cost of integration. FFX offers two thermostats for use in conjunction with the integrators: the Berendsen[117] and Bussi-Donadio-Parrinello[118] thermostats. The Berendsen thermostat causes the system temperature to decay exponentially toward a target temperature, but it does not produce particle velocities that are consistent with sampling from the canonical ensemble. The Bussi thermostat can be considered a global version of the (local) Langevin thermostat described below, and it produces particle velocities consistent with rigorous sampling from the canonical ensemble.

The Langevin dynamics[119,120] integrator incorporates two forces: a viscous damping force proportional to particle velocity and a random force representing the effects of collisions with molecules in the environment. The random forces are pulled from a Gaussian distribution with a mean of zero. The magnitude of both the viscous damping and random forces are controlled by

a collision frequency parameter (default in FFX is 91/psec to mimic water-like viscosity[121]). In this way, the Langevin integrator provides rigorous temperature control. Finally, pressure control is achieved through a Monte Carlo barostat that obeys Lattice system constraints[122].

Implementations. The default implementation for performing molecular dynamics leverages shared memory parallelism constructs defined by the Parallel Java API. This code path is currently recommended for small systems (*e.g.*, pharmaceutical crystals), for use with orthogonal space tempering, or for refinement against experimental data. For GPU acceleration of systems without space group symmetry, FFX offers an interface with OpenMM[123]. This latter code path is recommended for simulating large systems such as proteins and/or DNA solvated in a periodic box of water.

Constant pH Molecular Dynamics. FFX comes packaged with the first continuous constant pH molecular dynamics (CpHMD) algorithm for a polarizable atomic multipole potential[32]. The dynamics of the constant pH method are governed by an extended Hamiltonian of the form[124,125]

$$\mathcal{H}(\mathbf{X},\boldsymbol{\theta}) = U_{bond}(\mathbf{X}) + U_{nbond}(\mathbf{X},\boldsymbol{\theta}) + U^{*}(\boldsymbol{\theta}) + \sum_{i}^{N} 0.5 m_i \dot{X}_i^{\,2} + \sum_{k}^{N_{titr}} 0.5 m_k \dot{\theta}_k^{\,2}$$

Equation 25

where *i* is the index over atoms and *k* is the index over both titration and tautomer extended system variables. $\mathbf{X}$ is the Cartesian coordinate vector, and $\boldsymbol{\theta}$ is a vector of both titration and tautomer extended system particles. The titration (λ_k) and tautomer (ζ_k) states are bound between 0 and 1 through the relation, $\lambda_k \; or \; \zeta_k = sin^2(\theta_k)$. These particles are handled by a Langevin integrator with custom friction and particle mass parameters optimized for the algorithm. The θ particles propagate along with atomic particles, although calculation of the extended system forces is currently restricted to a CPU-only implementation. However, there is support for a hybrid CPU/GPU approach to accelerate the method. To achieve the acceleration, θ particles are "frozen" while the atomic coordinates are propagated on the GPU allowing coordinate/titration

steps to be followed by a defined number of coordinate-only steps to achieve improved sampling[32]. Sampling is further enhanced through use of pH replica exchange (RepEx)[126]. The CpHMD RepEx protocol runs multiple simulations on a pH ladder simultaneously. Throughout the simulations, there are periodic attempts to exchange pH's between simulations according to the Metropolis criterion.

$$P = \begin{cases} 1 & if\ \Delta \leq 0 \\ e^{-\beta \Delta E} & otherwise \end{cases}$$

Equation 26

Where β is given by $1/k_B T$ and ΔE is defined as

$$\Delta E = U_{pH}(\mathbf{X}_{A\ at\ B}, \boldsymbol{\theta}_{A\ at\ B}) + U_{pH}(\mathbf{X}_{B\ at\ A}, \boldsymbol{\theta}_{B\ at\ A}) - \left(U_{pH}(\mathbf{X}_A, \boldsymbol{\theta}_A) + U_{pH}(\mathbf{X}_B, \boldsymbol{\theta}_B)\right)$$

Equation 27

where A and B represent the two ensembles considered in the exchange. The pH exchange protocol also supports the CPU/GPU hybrid acceleration method described above. The combination of replica exchange and GPU acceleration results in the first tractable CpHMD simulations of proteins using the AMOEBA force field.

Free Energy Calculations

The *Thermodynamics* command supports estimation of free energy differences using two complimentary approaches. The first approach applies free energy perturbation (FEP), the Bennett Acceptance Ratio (BAR) method, and/or Multi-State BAR (MBAR) to sample from ensembles defined by pairs of λ values. It then accumulates the overall free energy difference (and its statistical uncertainty) over a path defined by a series of λ windows[127,128]. The second approach uses the orthogonal space tempering (OST) biased sampling strategy to estimate the free energy difference while overcoming hidden barriers[39,129]. Future work will access opportunities to re-weight biased samples from OST and thereby permit evaluation with MBAR.

Free Energy Perturbation, BAR and MBAR. For the BAR and MBAR methods, the λ values range from 0 to 1, where λ=0 indicates the initial state and λ=1 indicates the end state. Intermediate simulation windows sample from an ensemble that represents an unphysical alchemical state. The resulting samples are stored in an archive file (.arc) for each λ value, which can then read by the *BAR* or *MBAR* commands during estimation of free energy differences.

For example*, Thermodynamics* can be used to estimate the free energy difference associated with amino acid substitutions within protein structure. In this case, the thermodynamic path is broken into four sets of simulations to slowly remove the wildtype (WT) amino acid and slowly grow in the mutant (MT) amino acid. The first set turns off the electrostatics from the WT side chain, the second turns off van der Waals from the WT side chain, the third turns on the van der Waals for the MT side chain, and the final simulation turns on the electrostatics for the MT side chain.

To remove the electrostatics contribution of the WT side chain, FFX splits the simulation into a set number of alchemical intermediate simulations, typically referred to as windows. The user sets the alchemical atoms to be affected by the λ states (*i.e.,* the side chain atoms). Each window will have a different λ value evenly distributed across *n* windows from zero to one. The λ value acts as a dial to tune the contribution from the electrostatics from on (λ=1) to off (λ=0). FFX initiates windowed alchemical simulations with *Thermodynamics* when *n* windows are set. Each simulation is run in parallel and has GPU acceleration. The result is *n* archive files with snapshots from thermodynamics simulations used for further analysis to estimate the overall free energy change. The *BAR* command estimates the free energy difference from the collected samples at each λ value[127].

An extension of BAR is MBAR where equilibrium samples from multiple thermodynamic states can be utilized to construct a statistically optimal free energy estimator. MBAR reduces to BAR when used on two thermodynamic states.

Our recent expansion of the dual topology framework to directly estimate free energy differences between polymorphs are presented as an additional demonstration for both BAR and MBAR. A dual topology simulation via the *Thermodynamics* command was used to sample states between experimental crystals with lattice parameters and coordinates minimized to a convergence criteria of 0.05 kcal/mol/Å according to parameters generated by PolType2[130] for 2-((4-(2-(3,4-dichlorophenyl)ethyl)phenyl)amino)benzoic acid (CSD ID: XAFPAY) also known as compound XXIII. Symmetry operators to map between the two polymorph end states were generated via the PAC algorithm. Individual symmetry operators were generated for each of the aromatic rings and their constituents as shown in Figure 5. Several hydrogen atoms had inter-polymorph distances larger than 1.0 Å after the application of the symmetry operator that was applied to match their bonded heavy atom between polymorphs, therefore those hydrogen atoms were treated as alchemical.

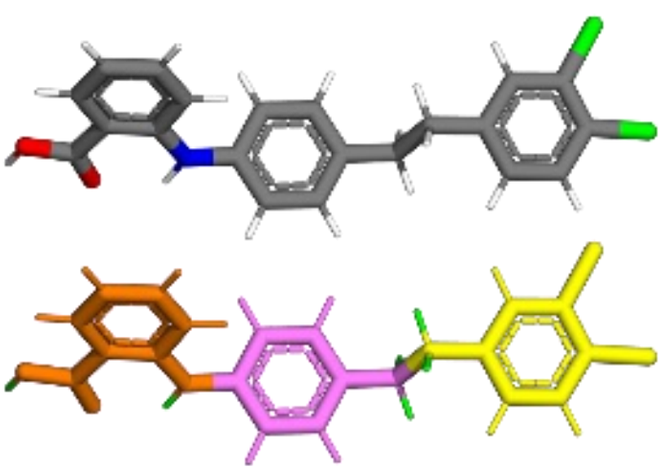

Figure 5. The upper panel shows compound XXIII colored by atomic number with chlorine green, oxygen red, nitrogen blue, carbon grey, and hydrogen white. The lower panel is colored by dual topology groups, with orange, violet, and yellow atoms each assigned a unique symmetry operator to map between polymorphs. The hydrogen atoms colored green in the lower panel were given independent degrees of freedom for the two topologies.

One CPU core was used to generate sampling for each of 21 lambda windows in parallel. Each window sampled for one nanosecond (~11.1 hours on an Intel® Xeon® Gold 6330 CPU at 2.00 GHz) with coordinate snapshots saved each picosecond. The last 900 snapshots were

evaluated MBAR to computed polymorph relative free energy differences, with results given in Table *3* and Table *4*. For comparison, relative lattice potential energy differences are provided for AMOEBA (which was used for the free energy difference calculations) along with several other approaches featured in the 6th blind test of organic crystal structure prediction organized by the Cambridge Crystallographic Data Centre[131].

Table 3. Relative lattice energies for the experimental polymorphs of Compound XXIII using a variety of models, compared to AMOEBA relative polymorph free energy differences (kcal/mol).

Model[a]	Method	Form				
		A	B	C	D	E
Team 3: Day *et al.*	Multipoles and exp-6	0.3	1.3	0.0	0.6	0.1
Team 5: van Eijck	Charges and exp-6	1.0	0.0	1.3	1.3	1.1
Team 14: Neumann *et al.*	PBE + Neuman-Perrin	0.9	0.0	0.0	0.7	0.5
Team 18: Price *et al.*	Multipoles and exp-6	2.3	0.0	0.8	2.2	1.3
Potential Energy	AMOEBA	1.30	0.00	1.80	1.34	1.78
MBAR Relative ΔG	AMOEBA	0.39	0.00	0.87	0.66	0.84

[a]Model entries that start with "Team" were converted from Table S12 of the 6th blind test of organic crystal structure prediction.

The estimated free energy differences from both BAR and MBAR follow the trends observed from the AMOEBA lattice potential energy differences. Individual free energy differences between each of the experimental polymorphs, their associated uncertainties, and cycle closure errors can be found in Table 4.

Table 4. Free energy differences and uncertainties in kcal/mol per molecule between experimental polymorphs of compound XXIII computed using MBAR.

Transformation	ΔG	Uncertainty
B→D	0.656	0.067
D→E	0.183	0.097
E→C	0.027	0.092
C→A	-0.482	0.095
A→B	-0.386	0.065
Cycle Closure	-0.001	

Orthogonal Space Tempering. First order generalized ensemble (GE) methods (e.g., metadynamics[132]) can eliminate barriers along the chosen variable path (*i.e.*, λ). However, free energy barriers perpendicular to the variable path can impede exhaustive sampling of the entire free energy surface (*e.g.*, conformational barriers)[133]. FFX implements the 2nd order GE orthogonal space tempering (OST) sampling[8] to cross free energy barriers by combining transition-tempered metadynamics[134] with the orthogonal space random walk method[39]. The total potential energy of the OST ensemble, $U_{\mathrm{OST}}(\lambda, \mathbf{x})$, is the summation of AMOEBA force field energy terms and the sum of a 1-dimensional and 2-dimensional time-dependent bias, $f_m(\lambda) + g_m(\lambda, \mathrm{F}_\lambda)$ as demonstrated in Equation 28.

$$U_{\mathrm{OST}}(\lambda, \mathbf{x}) = U_{\mathrm{AMOEBA}}(\lambda, \mathbf{x}) + f_m(\lambda) + g_m(\lambda, \mathrm{F}_\lambda)$$

Equation 28

The 1-dimensional bias, $f_m(\lambda)$, is obtained through the thermodynamic integration as given by

$$f_m(\lambda) = -\int_0^{\lambda} \langle \partial U/\partial\lambda \rangle_{\acute{\lambda}} \mathrm{d}\acute{\lambda}$$

Equation 29

where the ensemble average is further clarified in Equation 30 where $\beta = \frac{1}{k_{\mathrm{B}}T}$.

$$\langle \partial U/\partial\lambda \rangle_{\acute{\lambda}} = \frac{\int_{\partial U/\partial\lambda} \partial U/\partial\lambda \cdot \exp[\beta g_m(\lambda, \mathrm{F}_\lambda)]\ \delta(\lambda - \acute{\lambda})}{\int_{\partial U/\partial\lambda} \exp[\beta g_m(\lambda, \mathrm{F}_\lambda)]\ \delta(\lambda - \acute{\lambda})}$$

Equation 30

The time-dependent 2-dimensional repulsive potentials, $g_m(\lambda, \mathrm{F}_\lambda)$, are defined in Equation 31.

$$g_m(\lambda, F_\lambda) = \sum_{t_i} h(t_i) \cdot \exp\left[\frac{|\lambda - \lambda(t_i)|^2}{2{w_1}^2} \cdot \frac{|F_\lambda - F_\lambda(t_i)|^2}{2{w_2}^2}\right]$$

Equation 31

The "tempering" in OST denotes a non-constant bias height, $h(t_i)$, that decreases as the simulation proceeds. The intention of the bias is to progressively flatten the path. Therefore, the bias height decreases asymptotically towards 0 based on the following expression.

$$h(t_i) = h(t_0) \cdot \exp\left[\frac{\min\left(0, V_{\text{th}} - V^*(t_i)\right)}{\Delta T}\right]$$

Equation 32

where the initial Gaussian bias height $h(t_0)$ is typically chosen as ~0.02-0.05 kcal/mol with the standard deviation equal to two bins in either dimension ($w_1 = 0.01$ and $w_2 = 4.0$ kcal/mol). The height is also truncated after five bins during evaluation of the 2D bias. Tempering begins after a small amount of bias has been added along the entire path defined by the tempering threshold (V_{th}). The rate of exponential decay is determined by the ΔT parameter that has a default value of $2 \cdot k_B T$. Finally, the amount of decay is tempered based on $V^*(t_i)$ defined in Equation 33 where the max operation for each fixed λ is over the range of F_λ values (*i.e.*, the 2D g_m histogram is reduced to a 1D function of λ), followed by the min operation over λ.

$$V^*(t_i) = \min_{\lambda}\left[\max_{F_\lambda} g_m(\lambda, F_\lambda)\right]$$

Equation 33

Figure *6* presents a visualization of the OST algorithm as applied to a simulation of carbamazepine as it transitions between vacuum to crystalline phases. In Figure *6*A, the ensemble average partial derivative of the potential energy with respect to the path variable, λ, is plotted as a solid black line. The deposition free energy difference at each λ relative to the vacuum state (*i.e.*, $\lambda = 0$) is represented as a dashed blue line. Figure *6*B demonstrates a contour plot of the total OST bias (*i.e.*, the summation of both 1-D and 2-D bias contributions). The addition of OST bias promotes sampling away from the free energy minimum while in the crystalline phase

(near or at $\lambda = 1$), encouraging the exploration of free energy barriers along $\partial U/\partial\lambda$ (orthogonal to the λ path). Figure 6C features only the 2-dimensional bias contribution as a contour plot by removing the 1D bias which is constant for each λ value. The partial derivatives needed for OST ($\partial U/\partial X$, $\partial U/\partial\lambda$, $\partial^2 U/\partial\lambda^2$, $\partial^2 U/\partial X\partial\lambda$) are supported for softcore van der Waals interactions, softcore charge (or multipolar) interactions, and for polarization energy contributions within the CPU code path[6], but not yet available within OpenMM[129].

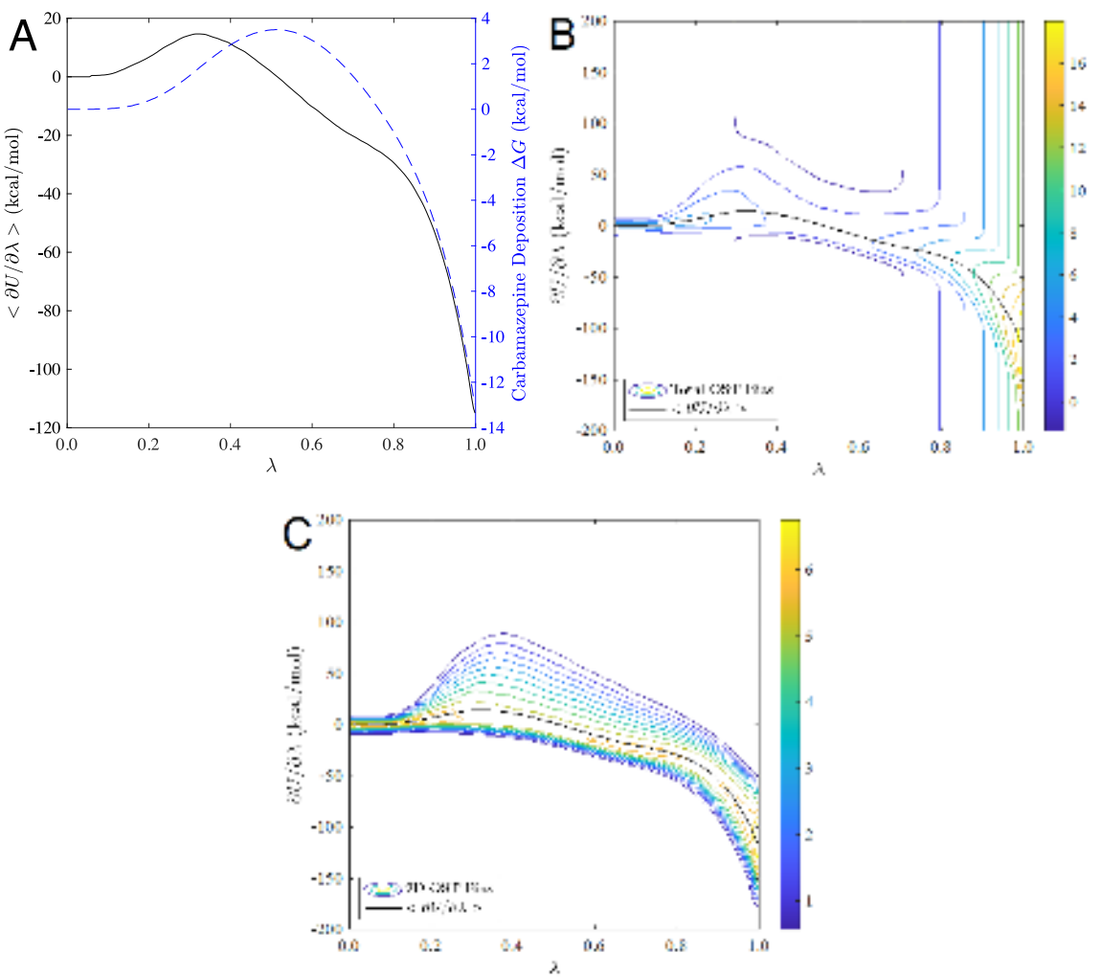

Figure 6. Plots illustrating the use of the OST sampling approach on a carbamazepine simulation between vacuum ($\lambda = 0$) and crystalline ($\lambda = 1$) phases. Panel A represents the ensemble average partial derivative of the potential energy with respect to the path variable λ, $\langle\partial U/\partial\lambda\rangle$, (solid black line) and the deposition free energy difference obtained as the integration over the phase transition path (dashed blue line). Panel B is a contour plot of the combined contributions from the 1D and 2D biases whereas panel C is a contour plot of only the 2D bias contributions. Both panels B and C show $\langle\partial U/\partial\lambda\rangle$ as a function of λ.

Properties and Analysis

A family of commands leverage coordinate snapshots and trajectories to analyze the properties of organic materials and biomolecules. The *Energy* command provides a summary of the potential energy contributions for all terms in use (*e.g.*, bonds, angles, torsion, van der Waals, permanent electrostatics, polarization, and various restraints). The *Volume* command supports calculation of molecular volume and surface area using either the GaussVol[135] or the Connolly algorithm[136]. The former is limited to calculation of van der Waals volume and surface area. The latter additionally supports calculation of both 1) molecular volume and surface area and 2) solvent excluded volume and solvent accessible surface area.

BAR and *MBAR* can be applied to single or dual topology systems after simulation with molecular dynamics or Metropolis Monte Carlo techniques[137,138]. *BAR* can save Tinker .bar files that contain energy evaluations for snapshots in each λ value, which is particularly useful for re-evaluating free energy differences using subsets of the coordinate snapshots (*e.g.,* to compare free energy difference estimates from the first and second half of trajectories). With an additional script *SortArc*, coordinate snapshots from thermodynamics simulations using the replica exchange algorithm can be reorganized and analyzed with BAR[139]. Specifically, *SortArc* sorts the multi-state snapshot files into archive files containing snapshots for only a single state (*i.e.,* all snapshots at λ=1 will be contained in a single .arc file).

Finally, the *Histogram* command is used to load a 2D count matrix from an OST simulation (stored in a histogram file *.his). *Histogram* then leverages the counts to first compute the ensemble average partial derivative of the potential energy with respect to the state variable $\langle \partial U/\partial \lambda \rangle$ followed by estimation of the free energy difference as a function λ using thermodynamic integration. The *Histogram* command can optionally save out text files that are input to a simple Matlab script that plots the 1D potential of mean force (PMF) and 2D OST bias to visualize the free energy surface as a function of both λ and $\partial U/\partial \lambda$.

Rotamer Optimization of Many-Body Potential

Rotamer optimization can be performed on a protein using the *ManyBody* command. The total energy of the protein is optimized with the many-body expansion through movement of defined side chain positions relative to the backbone (rotamers). The user can choose between two rotamer libraries with rotamers for every amino acid including the protonated and deprotonated form of titratable residues[140,141]. *ManyBody* results in an optimized structure with side chains in the global minimum energy conformation (GMEC). The user can improve the rotamer optimization by removing default approximations within the many-body expansion. The distance cutoffs for both two and three body energy terms can be increased to capture interactions of more distal residues. Users can adjust energy cutoffs for clash pruning, add or remove the three-body energy term, or allow soft-coring of clashes. There are multiple residue selection algorithms that include providing start and end residues as well as a user-specified list of residues. Rotamer optimization with *ManyBody* improves the overall structural quality of proteins.

For example, rotamer optimization with local minimization was applied to AlphaFold2[142] deep learning algorithm predicted isoform-specific protein structures for 218 protein-coding genes found in the Deafness Variation Database (DVD)[143]. MolProbity algorithm evaluated structures before and after optimization to quantify the improvement in atomic clashes, backbone angles, and side chain conformations[5]. Structures from AlphaFold2 had an average clash score of 20.75 (*i.e.*, number of unphysical overlaps per 1000 atoms), and the overall MolProbity score was 2.86 (*i.e.*, the protein models were consistent with those from 2.86 Å resolution diffraction data). After optimization with FFX, the average clash score of the protein models dropped to only 0.11 and, the protein structures had an average MolProbity score of 0.97, which is consistent with models from atomic resolution diffraction data.

Refinement Against X-ray, Neutron & Joint X-ray/Neutron Targets

FFX serves as platform to systematically explore the use of advanced potential energy functions (*e.g.*, AMOEBA) coupled to rigorous long-range electrostatics (*e.g.*, PME) in the context of both local and global optimization strategies during refinement of biomolecular models against experimental data sets. The target function weights the contributions of the data (*e.g.*, X-ray diffraction) with prior chemical knowledge. This can be motivated within a Bayesian framework where the probability of the model parameters $\mathbf{X}$ (*e.g.*, atomic coordinates, b-factors, occupancies) given the data **D** is given by $p(\mathbf{X}|\mathbf{D})$ and is proportional to

$$p(\mathbf{X}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{X})e^{(-U(\mathbf{X})/k_BT)}$$

Equation 34

where $U(\mathbf{X})$ is a force field potential energy, k_B is Boltzmann's constant, T is absolute temperature, and $p(\mathbf{D}|\mathbf{X})$ is the likelihood of the data given the model. For convenience, the maximum of $p(\mathbf{X}|\mathbf{D})$ can be found by minimization of the negative logarithm of Equation 34 as given by

$$E(\mathbf{X}) = -w_A \ln[p(\mathbf{D}|\mathbf{X})] + U(\mathbf{X})$$

Equation 35

where the weight of the data $w_A = k_BT$ should be approximately 0.6 near 300 K[10]. FFX supports reciprocal space refinement against either X-ray or neutron diffraction data, and joint X-ray plus neutron refinement using the target function

$$E(\mathbf{X}) = -w_{\mathrm{A,Xray}} \ln\left[p\left(\mathbf{D}_{\mathrm{Xray}}|\mathbf{X}\right)\right] - w_{\mathrm{A,Neutron}} \ln[p(\mathbf{D}_{\mathrm{Neutron}}|\mathbf{X})] + U(\mathbf{X})$$

Equation 36

where the relative weight of each data source ($w_{\mathrm{A,Xray}}$, $w_{\mathrm{A,Neutron}}$) can be configured.

Real space refinement is also supported using CryoEM electron density maps or those from a Fourier synthesis such as $(F_o - F_c)$ or $(2F_o - F_c)$ where F_o are observed structure factors and F_c are calculated structure factors. Prior work has shown that PME electrostatics, especially when combined with a polarizable multipole force field, improves structure quality[1]. Further model improvements are afforded by use of a differentiable bulk solvent model[9] and global optimization

of sidechain rotamers using a many-body expansion of Equation 35 coupled to Goldstein elimination criteria[11]. Work remains to implement a convergent and efficient implementation of generalized Kirkwood continuum electrostatics under periodic boundary conditions, which will open the door to estimation of free energy differences between model conformations and chemical compositions.

Unit Testing and Continuous Integration

FFX currently includes more than 500 JUnit tests, with many building on the commands described previously (*e.g.*, *Energy*, *Thermodynamics*, *ManyBody*, *etc.*) to validate both core functionality and command line flags. The number of JUnit tests for each package and current percentage of code covered by each test is summarized in Table 5. No JUnit tests specific to Parallel Java have been created due to the careful work by its original author[60]. We also lack JUnit tests for the graphical user interface. The OpenMM classes are covered by JUnit tests in the Potential and Algorithms packages. A Jenkins continuous integration server is used to automatically run all FFX unit tests and generate the Force Field X website based on polling the GitHub repository. The FFX website includes documentation for all commands and properties generated from annotations within Java code. Future work will focus on expanding test coverage for emerging methods within the Algorithms and Refinement packages.

Table 5. The number of unit tests and code coverage for most of the FFX packages.

Package	Number of Tests	Code Coverage (%)
Algorithms	100	32%
Crystal	13	91%
Numerics	175	82%
OpenMM	-	58%
Potential	201	60%
Refinement	27	40%
Utilities	2	14%

Shared Memory, MPI and GPU Parallelization

Shared Memory Parallelization. FFX leverages Parallel Java (PJ) for shared memory parallelism (SMP)[60] based on classes that offer functionality that is analogous to OpenMP style pragmas. Rather than annotating a loop with a pragma, PJ defines various "ForLoop" classes that are extended and then executed by a collection of threads called a "ParallelTeam"[61]. By default, nonbonded forces are calculated using all available threads. For PME, the direct space and reciprocal terms can be calculated concurrently. The direct space contributions are organized using neighbor lists to distribute and balance the workload. Parallelization of the 3D convolution operation that is the basis of reciprocal space PME has been described previously[1].

Message Passing Across Nodes. FFX executes on a single process by default but supports the cooperation of multiple processes through use of the PJ *Scheduler* and MPI implementation of PJ. The *Scheduler* organizes execution of a parallel job across a cluster of multiple processes with a user defined number of threads per process and memory per process. The *Scheduler* defaults to evenly splitting all available cores if this property remains unspecified by the user. The *Scheduler* and MPI constructs can be used in conjunction with both the SMP approach described above and the GPU support described below.

For example, OST free energy difference calculations can be accelerated through MPI parallelization over multiple walkers that each contribute counts to the same 2D histogram (Multiple Walker OST). The *Scheduler* launches individual trajectories that each start from an identical value of the state variable λ or be distributed across the thermodynamic path (*i.e.*, each walker has a unique starting value of λ). The samples from any given walker are communicated to all other walkers such that each process is applying the same OST bias and thereby providing the same estimated free energy difference. The addition of walkers enhances sampling and facilitates convergence. MPI approaches are also employed in the context of both many-body optimization and replica exchange constant pH MD.

Usage of GPUs via OpenMM. OpenMM binaries are bundled with FFX to allow usage of GPUs[20]. Alternatively, source code can be built and used via JNA or JExtract. FFX has java implementations for the majority of C++ classes available in OpenMM (*i.e.,* Context, Platform, State, *etc.*). Many-Body Optimization employs OpenMM for force field energy calculations and in turn, for self, pair, and triple energy calculations. The initialization of Many-Body occurs on the CPU while the AMOEBA energy calculations are performed after creating an OpenMM context and moving to the GPU. The finalization of the global minimum energy conformation is passed back to the CPU[4]. GPU acceleration is also available for MD, AMOEBA/GK calculations, replica exchange constant pH MD (hybrid CPU-GPU implementation), *etc*. SMP and MPI can be used in conjunction with OpenMM.

Benchmarks

Energy and force timings for simulating carbamazepine crystalline units are presented in Table 6 for three polymorphs to showcase the efficiency gained by simulating asymmetric units relative to replicated unit cells. Furthermore, the carbamazepine polymorph deposition simulation that produced the plots in Figure *6* was performed utilizing two threads of a recent Intel® CPU (a Xeon® Gold 6330 CPU at 2.00 GHz). Simulating for 3.6 hours produced 1 nanosecond of sampling using the AMOEBA force field or more than 350 nanoseconds/day using all 112 threads / 56 cores of a dual CPU configuration.

Table 6. Comparison of the asymmetric unit (ASU), unit cell (UC), replicated unit cell that satisfies minimum image convention for several experimental polymorphs of carbamazepine.

CSD ID	Crystalline Unit	Space Group	Number of Molecules	Degrees of Freedom	Time for Energy and Force Evaluations (s)
CBMZPN 02	ASU	$P2_1/n$	1	90	0.027
	UC	$P1$	4	360	0.029
	4x3x3 UC	$P1$	144	12,960	0.733
CBMZPN 12	ASU	$C2/c$	1	90	0.026
	UC	$P1$	8	720	0.055
	2x5x3 UC	$P1$	240	21,600	1.116
CBMZPN 16	ASU	$Pbca$	1	90	0.034
	UC	$P1$	8	720	0.046
	4x3x2 UC	$P1$	192	17,280	0.894

Rotamer optimization was performed for turkey-ovomucoid third domain, a 56 amino acid protein structure requiring ~112.5 thousand AMOEBA/GK energy calculations with zero, one, two, and four GPUs on Nvidia A10s with 28 Intel CPU cores per GPU. The simulation experienced an 9.4X speed up from 7.6 AMOEBA energy calculations per second to 71.7 per second when increasing from no GPU to four GPUs.

Table 7. AMOEBA/GK energy evaluations per second with different numbers of GPU's performing a ManyBody optimization on a 56-residue turkey-ovomucoid third domain protein.

Architecture	AMOEBA Energy Evaluations / sec
CPU	7.6
1 GPU	29.9
2 GPUs	48.8
4 GPUs	71.7

Finally, for molecular dynamics simulations that can be offloaded to OpenMM (*e.g.*, unit cell lengths that are larger than twice the non-bonded cut-off and do not require symmetry operators), the benchmarks described by the OpenMM developers[20,144] are representative of the performance in FFX.

Conclusions

This article has demonstrated significant use cases and advancements available in FFX for atomic resolution modeling of organic materials and biomolecules. Specifically, we have highlighted FFX's handling of crystal structures and data, the generalized Kirkwood implicit solvent model, constant-pH MD for the polarizable AMOEBA force field, dual topology methods, and global side chain optimization. FFX development will continue with novel algorithms and advanced treatment of force fields, acid-base chemistry, and prediction of crystal properties. Some methods under active development include a statistical mechanics method for accelerated acid-base chemistry calculations, AMOEBA folding and binding free energy difference predictions for amino acid mutations, GPU acceleration of constant pH MD, and methods for relative

polymorph free energy differences. It is our hope the open-source and freely available FFX software can serve as a computational microscope to understand the biophysics of organic materials and biomolecules.

Acknowledgements

Authors ACT, RACG and MRT were supported by the NSF (National Science Foundation) Graduate Research Fellowship under Grant No. 000390183. RACG and AJN were partially supported by a Ballard and Seashore Dissertation Fellowship from the University of Iowa. RACG, GQ, and LMC were partially supported by fellowships from the University of Iowa Office of Undergraduate Research. Authors JWP and PR were supported by NIH grants R01GM114237 and R01GM106137. Author JKS was supported by NIH grant R35GM148261. Author MJS NSF grant CHE-1751688. Authors MJS and RJHS were supported by NIH grant R01DC012049. Authors MJS and JJM were supported by Simons Foundation SFARI Award 1019623.

Author Declarations

J.-P. Piquemal, J. Ponder, and P. Ren are cofounders of Qubit Pharmaceuticals.

Data Availability

The Force Field X code is available on GitHub (https://github.com/SchniedersLab/forcefieldx) under the GNU General Public License, version 3, with the Classpath Exception. Force Field X documentation, including instructions to install precompiled versions or build the software from source code, is available from the U. of Iowa (https://ffx.biochem.uiowa.edu).

References

1 Schnieders, M. J., Fenn, T. D. & Pande, V. S. Polarizable atomic multipole X-ray refinement: Particle mesh Ewald electrostatics for macromolecular crystals. *J. Chem. Theory Comput.* **7**, 1141-1156 (2011). https://doi.org/10.1021/ct100506d

2 Corrigan, R. A. *et al.* Implicit Solvents for the Polarizable Atomic Multipole AMOEBA Force Field. *J. Chem. Theory Comput.* **17**, 2323-2341 (2021). https://doi.org/10.1021/acs.jctc.0c01286

3 Corrigan, R. A. *et al.* A generalized Kirkwood implicit solvent for the polarizable AMOEBA protein model. *The Journal of Chemical Physics* **159** (2023). https://doi.org/10.1063/5.0158914

4 Tollefson, M. R. *et al.* Structural insights into hearing loss genetics from polarizable protein repacking. *Biophys. J.* **117**, 602-612 (2019). https://doi.org/10.1016/j.bpj.2019.06.030

5 Tollefson, M. R. *et al.* Assessing variants of uncertain significance implicated in hearing loss using a comprehensive deafness proteome. *Hum. Genet.* **142**, 819-834 (2023). https://doi.org/10.1007/s00439-023-02559-9

6 Schnieders, M. J. *et al.* The structure, thermodynamics, and solubility of organic crystals from simulation with a polarizable force field. *J. Chem. Theory Comput.* **8**, 1721-1736 (2012). https://doi.org/10.1021/ct300035u

7 Nessler, I. J., Litman, J. M. & Schnieders, M. J. Toward polarizable AMOEBA thermodynamics at fixed charge efficiency using a dual force field approach: application to organic crystals. *Phys. Chem. Chem. Phys.* **18**, 30313-30322 (2016). https://doi.org/10.1039/C6CP02595A

8 Litman, J., Thiel, A. C. & Schnieders, M. J. Scalable indirect free energy method applied to divalent cation-metalloprotein binding. *J. Chem. Theory Comput.* **15**, 4602-4614 (2019). https://doi.org/10.1021/acs.jctc.9b00147

9 Fenn, T. D., Schnieders, M. J. & Brunger, A. T. A smooth and differentiable bulk-solvent model for macromolecular diffraction. *Acta Crystallogr. D* **66**, 1024-1031 (2010). https://doi.org/10.1107/S0907444910031045

10 Fenn, T. D. & Schnieders, M. J. Polarizable atomic multipole X-ray refinement: Weighting schemes for macromolecular diffraction. *Acta Crystallogr. D* **67**, 957-965 (2011). https://doi.org/10.1107/S0907444911039060

11 LuCore, Stephen D. *et al.* Dead-end elimination with a polarizable force field repacks PCNA structures. *Biophys. J.* **109**, 816-826 (2015). https://doi.org/http://dx.doi.org/10.1016/j.bpj.2015.06.062

12 Ren, P. & Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **107**, 5933-5947 (2003).

13 Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **114**, 2549-2564 (2010). https://doi.org/10.1021/jp910674d

14 Ren, P., Wu, C. & Ponder, J. W. Polarizable atomic multipole-based molecular mechanics for organic molecules. *Journal of Chemical Theory and Computation* **7**, 3143-3161 (2011).

15 Shi, Y. *et al.* Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* **9**, 4046-4063 (2013). https://doi.org/10.1021/ct4003702

16 Zhang, C. *et al.* AMOEBA polarizable atomic multipole force field for nucleic acids. *J. Chem. Theory Comput.* **14**, 2084-2108 (2018). https://doi.org/10.1021/acs.jctc.7b01169

17 Harger, M. *et al.* Tinker-OpenMM: Absolute and Relative Alchemical Free Energies using AMOEBA on GPUs. *J. Comput. Chem.* **38**, 2047-2055 (2017). https://doi.org/10.1002/jcc.24853

18 Rackers, J. A. *et al.* Tinker 8: Software Tools for Molecular Design. *Journal of Chemical Theory and Computation* **14**, 5273-5289 (2018). https://doi.org/10.1021/acs.jctc.8b00529

19 Lagardère, L. *et al.* Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chemical Science* **9**, 956-972 (2018). https://doi.org/10.1039/C7SC04531J

20 Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, 17 (2017). https://doi.org/10.1371/journal.pcbi.1005659

21 Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography* **54**, 905-921 (1998).

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

23 Gosling, J. & McGilton, H. The Java language environment. *Sun Microsystems Computer Company* **2550**, 38 (1995).

24 Gosling, J. *The Java language specification*. (Addison-Wesley Professional, 2000).

25 Arnold, K., Gosling, J. & Holmes, D. *The Java programming language*. (Addison Wesley Professional, 2005).

26 Würthinger, T. *et al.* in *Proceedings of the 2013 ACM international symposium on New ideas, new paradigms, and reflections on programming & software.* 187-204.

27 Oracle. *Build faster, smaller, leaner applications*, <https://www.graalvm.org> (2023).

28 Clarkson, J. *et al.* in *Proceedings of the 15th International Conference on Managed Languages & Runtimes* Article 4 (Association for Computing Machinery, Linz, Austria, 2018).

29 Fumero, J. *et al.* in *Proceedings of the 15th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments* 165–178 (Association for Computing Machinery, Providence, RI, USA, 2019).

30 Zhang, Q., Xu, L. & Xu, B. Python meets JIT compilers: A simple implementation and a comparative evaluation. *Software: Practice and Experience* **n/a** https://doi.org/https://doi.org/10.1002/spe.3267

31 Oracle. *High-performance modern Python*, <https://www.graalvm.org/python> (2023).

32 Thiel, A. C. *et al.* Constant-pH Simulations with the Polarizable Atomic Multipole AMOEBA Force Field. *J. Chem. Theory Comput.* **20**, 2921-2933 (2024). https://doi.org/10.1021/acs.jctc.3c01180

33 Kirkwood, J. G. Theory of solutions of molecules containing widely separated charges with special application to zwitterions. *J. Chem. Phys.* **2**, 351-361 (1934).

34 Schnieders, M. J. & Ponder, J. W. Polarizable atomic multipole solutes in a generalized Kirkwood continuum. *J. Chem. Theory Comput.* **3**, 2083-2097 (2007). https://doi.org/10.1021/ct7001336

35 Corrigan, R. A. *et al.* Implicit Solvents for the Polarizable Atomic Multipole AMOEBA Force Field. *J Chem Theory Comput* **17**, 2323-2341 (2021). https://doi.org/10.1021/acs.jctc.0c01286

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

37 Essmann, U. *et al.* A smooth particle-mesh Ewald method. *J. Chem. Phys.* **103**, 8577-8593 (1995).

38 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.* **120**, 73-87 (2004).

39 Zheng, L., Chen, M. & Yang, W. Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. *Proceedings of the National Academy of Sciences* **105**, 20227-20232 (2008). https://doi.org/10.1073/pnas.0810631106

40 Min, D. *et al.* Practically Efficient QM/MM Alchemical Free Energy Simulations: The Orthogonal Space Random Walk Strategy. *Journal of Chemical Theory and Computation* **6**, 2253-2266 (2010). https://doi.org/10.1021/ct100033s

41 Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542 (1992).

42 Goldstein, R. F. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys J* **66**, 1335-1340 (1994). https://doi.org/10.1016/S0006-3495(94)80923-3

43 Tollefson, M. R. *et al.* Structural Insights into Hearing Loss Genetics from Polarizable Protein Repacking. *Biophysical Journal* **117**, 602-612 (2019). https://doi.org/10.1016/j.bpj.2019.06.030

44 Dybeck, E. C. *et al.* A Comparison of Methods for Computing Relative Anhydrous–Hydrate Stability with Molecular Simulation. *Cryst. Growth Des.* **23**, 142-167 (2023). https://doi.org/10.1021/acs.cgd.2c00832

45 Bu, F. *et al.* High-throughput genetic testing for thrombotic microangiopathies and C3 glomerulopathies. *J. Am. Soc. Nephrol.* **27**, 1245-1253 (2016). https://doi.org/10.1681/asn.2015040385

46 DeLuca, A. P. *et al.* Hypomorphic mutations in TRNT1 cause retinitis pigmentosa with erythrocytic microcytosis. *Hum. Mol. Genet.* **25**, 44-56 (2016). https://doi.org/10.1093/hmg/ddv446

47 Simpson, A. *et al.* LADD syndrome with glaucoma is caused by a novel gene. *Mol. Vis.* **23**, 179-184 (2017).

48 Lansdon, L. A. *et al.* The Use of Variant Maps to Explore Domain-Specific Mutations of FGFR1. *J. Dent. Res.* **96**, 1339-1345 (2017). https://doi.org/10.1177/0022034517726496

49 Schnieders, M. J. *et al.* A novel mutation (LEU396ARG) in OPA1 is associated with a severe phenotype in a large dominant optic atrophy pedigree. *Eye (Lond)* **32**, 843-845 (2018). https://doi.org/10.1038/eye.2017.303

50 Boese, E. A. *et al.* Novel Intragenic PAX6 Deletion in a Pedigree with Aniridia, Morbid Obesity, and Diabetes. *Curr. Eye Res.* **45**, 91-96 (2020). https://doi.org/10.1080/02713683.2019.1649704

51 Hagedorn, J. *et al.* Nanophthalmos patient with a THR518MET mutation in MYRF, a case report. *BMC Ophthalmol.* **20**, 388 (2020). https://doi.org/10.1186/s12886-020-01659-8

52 Bi, J. *et al.* Characterization of a TP53 Somatic Variant of Unknown Function From an Ovarian Cancer Patient Using Organoid Culture and Computational Modeling. *Clin. Obstet. Gynecol.* **63**, 109-119 (2020). https://doi.org/10.1097/grf.0000000000000516

53 Hinz, B. E. *et al.* In Silico and In Vivo Analysis of Amino Acid Substitutions That Cause Laminopathies. *Int J Mol Sci* **22** (2021). https://doi.org/10.3390/ijms222011226

54 Awotoye, W. *et al.* Whole-genome sequencing reveals de-novo mutations associated with nonsyndromic cleft lip/palate. *Sci Rep* **12**, 11743 (2022). https://doi.org/10.1038/s41598-022-15885-1

55 Tollefson, M. R. *et al.* Assessing variants of uncertain significance implicated in hearing loss using a comprehensive deafness proteome. *Hum Genet* (2023). https://doi.org/10.1007/s00439-023-02559-9

56 Boese, E. A. *et al.* GJA3 Genetic Variation and Autosomal Dominant Congenital Cataracts and Glaucoma Following Cataract Surgery. *JAMA Ophthalmol* (2023). https://doi.org/10.1001/jamaophthalmol.2023.3535

57 Nessler, A. J., Okada, O., Hermon, M. J., Nagata, H. & Schnieders, M. J. Progressive alignment of crystals: reproducible and efficient assessment of crystal structure similarity. *J. Appl. Crystallogr.* **55**, 1528-1537 (2022). https://doi.org/doi:10.1107/S1600576722009670

58 Nessler, A. J. *et al.* Crystal Polymorph Search in the NPT Ensemble via a Deposition/Sublimation Alchemical Path. *Cryst. Growth Des.* **24**, 3205-3217 (2024). https://doi.org/10.1021/acs.cgd.3c01358

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

60 Kaminsky, A. in *2007 IEEE International Parallel and Distributed Processing Symposium.* (IEEE).

61 Kaminsky, A. *Building Parallel Programs: SMPs, Clusters & Java*. (Course Technology Press, 2009).

62 Cooley, J. W. & Tukey, J. W. An Algorithm for the Machine Calculation of Complex Fourier Series. *Math. Comput.* **19**, 297-301 (1965). https://doi.org/10.2307/2003354

63 Temperton, C. Self-sorting mixed-radix fast Fourier transforms. *Journal of Computational Physics* **52**, 1-23 (1983). https://doi.org/https://doi.org/10.1016/0021-9991(83)90013-X

64 Challacombe, M., Schwegler, E. & Almlöf, J. in *Computational Chemistry: Reviews of Current Trends* Vol. Volume 1 *Computational Chemistry: Reviews of Current Trends* 53-107 (WORLD SCIENTIFIC, 1996).

65 Simmonett, A. C., Pickard, F. C., Schaefer, H. F. & Brooks, B. R. An efficient algorithm for multipole energies and derivatives based on spherical harmonics and extensions to particle mesh Ewald. *The Journal of Chemical Physics* **140**, 184101 (2014). https://doi.org/10.1063/1.4873920

66 Broyden, C. G. The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA J. Appl. Math* **6**, 76-90 (1970). https://doi.org/10.1093/imamat/6.1.76

67 Fletcher, R. A new approach to variable metric algorithms. *Computer J.* **13**, 317-322 (1970). https://doi.org/10.1093/comjnl/13.3.317

68 Goldfarb, D. A family of variable-metric methods derived by variational means. *Math. Comput.* **24**, 23-26 (1970).

69 Shanno, D. F. Conditioning of quasi-newton methods for function minimization. *Math. Comput.* **24**, 647-656 (1970).

70 de Boor, C. *A Practical Guide to Splines*. (Springer, 2001).

71 Cowtan, K. Generic representation and evaluation of properties as a function of position in reciprocal space. *J. Appl. Crystallogr.* **35**, 655-663 (2002). https://doi.org/doi:10.1107/S0021889802013420

72 Paszke, A. *et al.* in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* Article 721 (Curran Associates Inc., 2019).

73 Goldstein, R. F. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* **66**, 1335-1340 (1994). https://doi.org/http://dx.doi.org/10.1016/S0006-3495(94)80923-3

74 Beeman, D. Some multistep methods for use in molecular dynamics calculations. *Journal of Computational Physics* **20**, 130-139 (1976). https://doi.org/https://doi.org/10.1016/0021-9991(76)90059-0

75 Allen, M. P. Brownian dynamics simulation of a chemical reaction in solution. *Mol. Phys.* **40**, 1073-1087 (1980). https://doi.org/10.1080/00268978000102141

76 Guarnieri, F. & Still, W. C. A rapidly convergent simulation method: mixed Monte Carlo/stochastic dynamics. *J. Comput. Chem.* **15**, 1302–1310 (1994). https://doi.org/10.1002/jcc.540151111

77 Humphreys, D. D., Friesner, R. A. & Berne, B. J. A Multiple-Time-Step Molecular Dynamics Algorithm for Macromolecules. *The Journal of Physical Chemistry* **98**, 6885-6892 (1994). https://doi.org/10.1021/j100078a035

78 Qian, X. & Schlick, T. Efficient multiple-time-step integrators with distance-based force splitting for particle-mesh-Ewald molecular dynamics simulations. *The Journal of Chemical Physics* **116**, 5971-5983 (2002). https://doi.org/10.1063/1.1458542

79 Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. v., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690 (1984).

80 Bussi, G., Zykova-Timan, T. & Parrinello, M. Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *The Journal of Chemical Physics* **130** (2009). https://doi.org/10.1063/1.3073889

81 Frenkel, D. & Smit, B. *Understanding Molecular Simulation*. Second edn, 1-638 (Academic Press, 2002).

82 Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22**, 245-268 (1976). https://doi.org/10.1016/0021-9991(76)90078-4

83 Lee, S., Chen, M., Yang, W. & Richards, N. G. J. Sampling Long Time Scale Protein Motions: OSRW Simulation of Active Site Loop Conformational Free Energies in Formyl-CoA:Oxalate CoA Transferase. *J. Am. Chem. Soc.* **132**, 7252-7253 (2010). https://doi.org/10.1021/ja101446u

84 Park, J. *et al.* Absolute organic crystal thermodynamics: growth of the asymmetric unit into a crystal via alchemy. *J. Chem. Theory Comput.* **10**, 2781-2791 (2014). https://doi.org/10.1021/ct500180m

85 Sehnal, D. *et al.* BinaryCIF and CIFTools—Lightweight, efficient and extensible macromolecular data management. *PLoS Comput. Biol.* **16**, e1008247 (2020). https://doi.org/10.1371/journal.pcbi.1008247

86 Lafita, A. *et al.* BioJava 5: A community driven open-source bioinformatics library. *PLoS Comput Biol* **15**, e1006791 (2019). https://doi.org/10.1371/journal.pcbi.1006791

87 Steinbeck, C. *et al.* Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des* **12**, 2111-2120 (2006). https://doi.org/10.2174/138161206777585274

88 Fumero, J. *et al.* in *Proceedings of the 15th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments* (Association for Computing Machinery, 2019).

89 *picocli - a mighty tiny command line interface*, <https://picocli.info/> (2024).

90 Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225-11236 (1996). https://doi.org/10.1021/ja9621760

91 Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **105**, 6474-6487 (2001). https://doi.org/10.1021/jp003919d

92 Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **118**, 2309 (1996).

93 Kollman, P. *et al.* in *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications* (eds Wilfred F. van Gunsteren, Paul K. Weiner, & Anthony J. Wilkinson) 83-96 (Springer Netherlands, 1997).

94 Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **21**, 1049-1074 (2000). https://doi.org/https://doi.org/10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F

95 Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.* **65**, 712-725 (2006). https://doi.org/10.1002/prot.21123

96 MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586-3616 (1998).

97 Foloppe, N. & MacKerell, J., Alexander D. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **21**, 86-104 (2000). https://doi.org/https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2<86::AID-JCC2>3.0.CO;2-G

98 Mackerell Jr., A. D., Feig, M. & Brooks III, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25**, 1400-1415 (2004). https://doi.org/https://doi.org/10.1002/jcc.20065

99 Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127-6129 (1990).

100 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. *Journal of Physical Chemistry A* **101**, 3005-3014 (1997).

101 Onufriev, A., Case, D. A. & Bashford, D. Effective Born radii in the Generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **23**, 1297-1304 (2002).

102 Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246**, 122-129 (1995).

103 Grycuk, T. Deficiency of the Coulomb-field approximation in the Generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.* **119**, 4817-4826 (2003).

104 Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A. & Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *J. Chem. Theory Comput.* **3**, 156-169 (2007).

105 Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins: Struct., Funct., Bioinf.* **55**, 383-394 (2004).

106 Hudson, P. S. *et al.* Obtaining QM/MM binding free energies in the SAMPL8 drugs of abuse challenge: indirect approaches. *J. Comput. Aided Mol. Des.* **36**, 263-277 (2022). https://doi.org/10.1007/s10822-022-00443-8

107 Nocedal, J. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation* **35**, 773-782 (1980).

108 Liu, D. C. & Nocedal, J. On the Limited Memory Method for Large Scale Optimization. *Mathematical Programming B* **45**, 503-528 (1989).

109 Kirkpatrick, S., Gelatt, J. C. D. & Vecchi, M. P. Optimization by Simulated Annealing. *Science* **220**, 671-680 (1983). https://doi.org/10.1126/science.220.4598.67

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

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

112 Schnieders, M. J., Fenn, T. D., Pande, V. S. & Brunger, A. T. Polarizable atomic multipole X-ray refinement: application to peptide crystals. *Acta Crystallogr D Biol Crystallogr* **65**, 952-965 (2009). https://doi.org/10.1107/s0907444909022707

113 Lipparini, F. *et al.* Scalable Evaluation of Polarization Energy and Associated Forces in Polarizable Molecular Dynamics: I. Toward Massively Parallel Direct Space Computations. *J. Chem. Theory Comput.* **10**, 1638-1651 (2014). https://doi.org/10.1021/ct401096t

114 Verlet, L. Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical review* **159**, 98 (1967).

115 Swope, W. C., Andersen, H. C., Berens, P. H. & Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of chemical physics* **76**, 637-649 (1982).

116 Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible multiple time scale molecular dynamics. *The Journal of chemical physics* **97**, 1990-2001 (1992).

117 Berendsen, H. J., Postma, J. v., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **81**, 3684-3690 (1984).

118 Bussi, G. & Parrinello, M. Stochastic thermostats: comparison of local and global schemes. *Computer Physics Communications* **179**, 26-29 (2008).

119 Allen, M. Brownian dynamics simulation of a chemical reaction in solution. *Molecular Physics* **40**, 1073-1087 (1980).

120 Guarnieri, F. & Still, W. C. A rapidly convergent simulation method: Mixed Monte Carlo/stochastic dynamics. *Journal of Computational Chemistry* **15**, 1302-1310 (1994).

121 Shi, Y. Y., Wang, L. & Van Gunsteren, W. F. On the Approximation of Solvent Effects on the Conformation and Dynamics of Cyclosporin a by Stochastic Dynamics Simulation Techniques. *Mol Simulat* **1**, 369-383 (1988). https://doi.org/10.1080/08927028808080959

122 Frenkel, D. & Smit, B. *Understanding molecular simulation: from algorithms to applications*. (Academic Press San Diego, 2002).

123 Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).

124 Lee, M. S., Salsbury, F. R., Jr. & Brooks, C. L., 3rd. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **56**, 738-752 (2004). https://doi.org/10.1002/prot.20128

125 Khandogin, J. & Brooks, C. L., 3rd. Constant pH molecular dynamics with proton tautomerism. *Biophys J* **89**, 141-157 (2005). https://doi.org/10.1529/biophysj.105.061341

126 Wallace, J. A. & Shen, J. K. Continuous Constant pH Molecular Dynamics in Explicit Solvent with pH-Based Replica Exchange. *J Chem Theory Comput* **7**, 2617-2629 (2011). https://doi.org/10.1021/ct200146j

127 Bennett, C. H. Efficient Estimation of Free Energy Difference from Monte Carlo Data. *Journal of Computational Physics* **22**, 245-268 (1976).

128 Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 10 (2008). https://doi.org/10.1063/1.2978177

129 Schnieders, M. J. *et al.* The Structure, Thermodynamics and Solubility of Organic Crystals from Simulation with a Polarizable Force Field. *J Chem Theory Comput* **8**, 1721-1736 (2012). https://doi.org/10.1021/ct300035u

130 Walker, B., Liu, C., Wait, E. & Ren, P. Automation of AMOEBA polarizable force field for small molecules: Poltype 2. *J. Comput. Chem.* **43**, 1530-1542 (2022). https://doi.org/10.1002/jcc.26954

131 Reilly, A. M. *et al.* Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B* **72**, 439-459 (2016). https://doi.org/10.1107/S2052520616007447

132 Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *WIREs Computational Molecular Science* **1**, 826-843 (2011). https://doi.org/https://doi.org/10.1002/wcms.31

133 Zheng, L. & Yang, W. Practically efficient and robust free energy calculations: Double-integration orthogonal space tempering. *J. Chem. Theory Comput.* **8**, 810-823 (2012). https://doi.org/10.1021/ct200726v

134 Dama, J. F., Rotskoff, G., Parrinello, M. & Voth, G. A. Transition-Tempered Metadynamics: Robust, Convergent Metadynamics via On-the-Fly Transition Barrier Estimation. *J. Chem. Theory Comput.* **10**, 3626-3633 (2014). https://doi.org/10.1021/ct500441q

135 Zhang, B., Kilburg, D., Eastman, P., Pande, V. S. & Gallicchio, E. Efficient gaussian density formulation of volume and surface areas of macromolecules on graphical processing units. *J Comput Chem* **38**, 740-752 (2017). https://doi.org/10.1002/jcc.24745

136 Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709-713 (1983). https://doi.org/10.1126/science.6879170

137 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087-1092 (1953).

138 Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109 (1970). https://doi.org/https://doi.org/10.1093/biomet/57.1.97

139 Zhou, R. Replica exchange molecular dynamics method for protein folding simulation. *Methods Mol Biol* **350**, 205-223 (2007). https://doi.org/10.1385/1-59745-189-4:205

140 Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. The penultimate rotamer library. *Proteins* **40**, 389-408 (2000).

141 Ponder, J. W. & Richards, F. M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**, 775-791 (1987). https://doi.org/10.1016/0022-2836(87)90358-5

142 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org/10.1038/s41586-021-03819-2

143 Azaiez, H. *et al.* Genomic Landscape and Mutational Signatures of Deafness-Associated Genes. *Am J Hum Genet* **103**, 484-497 (2018). https://doi.org/10.1016/j.ajhg.2018.08.006

144 Eastman, P. *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B* **128**, 109-116 (2024). https://doi.org/10.1021/acs.jpcb.3c06662