Development and Comprehensive Benchmark of a High Quality AMBER-Consistent Small Molecule Force Field with Broad Chemical Space Coverage for Molecular Modeling and Free Energy Calculation

Bai Xue,[†,@] Qingyi Yang,[‡,@] Qiaochu Zhang,[†] Xiao Wan,[†] Dong Fang,[†] Xiaolu Lin,[†] Guangxu Sun,[†] Gianpaolo Gobbo,[¶] Fenglei Cao,[†] Alan M. Mathiowetz,[‡] Benjamin J. Burke,[§] Robert A. Kumpf,[§] Brajesh K. Rai,[‖] Geoffrey P.F. Wood,[⊥] Frank C. Pickard IV,[⊥] Junmei Wang,[#] Peiyu Zhang,[†] Jian Ma,[†] Yide Alan Jiang,[¶] Shuhao Wen,[†] Xinjun Hou,[*,‡] Junjie Zou,[*,†] and Mingjun Yang[*,†]

†*Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Floor 3, Sf Industrial Plant, No. 2 Hongliu Road, Fubao Community, Fubao Street, Futian District, Shenzhen 518045, China*

‡*Medicine Design, Pfizer Inc., 1 Portland Street, Cambridge, Massachusetts 02139, United States*

¶*XtalPi Inc., 245 Main Street, Cambridge, Massachusetts 02142, United States*

§*Medicine Design, Pfizer Inc., 10777 Science Center Drive, San Diego, CA 92121, United States*

‖*Machine Learning and Computational Sciences, Pfizer Inc., 610 Main Street, Cambridge, Massachusetts 02139, United States*

⊥*Pharmaceutical Science Small Molecule, Pfizer Inc., Eastern Point Road, Groton, Connecticut 06340, United States*

#*Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, United States*

@*These authors made equal contributions.*

E-mail: Xinjun.Hou@Pfizer.com; junjie.zou@xtalpi.com; mingjun.yang@xtalpi.com

Abstract

Biomolecular simulations have become an essential tool in contemporary drug discovery and molecular mechanics force fields constitute its cornerstone. Developing a

high quality and broad coverage general force field is a significant undertaking that requires substantial expert knowledge and computing resources, which is beyond the scope of general practitioners. Existing force fields originate from only a limited number of groups and organizations, and they either suffer from limited numbers of training sets, lower than desired quality because of oversimplified representations, or are costly for the molecular modeling community to access. To address these issues, in this work, we developed an AMBER-consistent small molecule force field with extensive chemical space coverage and we provide Open Access parameters for the entire modeling community. To validate our force field, we carried out benchmarks of quantum mechanics/molecular mechanics conformer comparison and free energy perturbation calculations on several benchmark data sets. Our force field achieves a higher level performance at reproducing quantum mechanics energies and geometries than two popular open-source force fields, OpenFF2 and GAFF2. In relative binding free energy calculations for 31 protein-ligand data sets, comprising 1079 pairs of ligands, the new force field achieves an overall root mean square error of 1.19 kcal/mol for $\Delta\Delta G$ and 0.92 kcal/mol for ΔG on a subset of 463 ligands without bespoke fitting to the data sets. The results are on par with the leading commercial series of OPLS force fields.

Introduction

Molecular mechanics force field (FF) parameters required to accurately model both biomacromolecules and small drug-like organic molecules has historically been one of the main topics of interests in molecular modeling and it has drawn even more attention in recent years with the ever increasing popularity of molecular simulations in drug discovery. One of the main challenges in FF development is the need to cover a large and diverse chemical space for drug-like molecules. Thanks to efforts from both the academic and industrial communities, several general small molecule FFs have been developed and widely applied in combination with the corresponding ones for biomacromolecules. Examples of general small molecule FFs include

but are not limited to GAFF[1] compatible with AMBER biomolecule FFs,[2] CGENFF[3] compatible with CHARMM biomolecule FFs,[4] a series of OPLS FFs[5,6] from either academic or industrial research groups compatible with the corresponding OPLS biomolecule FFs,[7,8] etc. These FFs adopt a similar functional form and comprise the FF parameters that include both bonded (bond, angle, and torsion) and non-bonded (electrostatic and vdW) terms. The bonded terms are usually parametrized over the optimized geometry, energy, or frequency from quantum mechanics (QM) calculations and fine-tuned with experimental crystal structures from the CCDC[9] database. The vdW terms are parametrized against experimentally measured liquid density and heat of vaporization data.[10] The atomic charges for electrostatic interactions are parametrized against either QM electrostatic potentials (ESPs)[11,12] or molecule-water interaction energy and distance.[13] In the later stages of parameter optimization, some FF development protocols may also include experimental measurements such as hydration free energies[14] or even protein-ligand binding affinity data[6] to fine-tune the non-bonded parameters. Recent advances in software tools have even made it possible to automatically adjust vdW parameters.[15] For a more detailed discussion of recent advances in FF development, we refer the readers to He et al.[16] Despite many decades of FF development efforts in academia and industry, access to more accurate FFs continues to be a challenge for most common practitioners in the molecular simulation community.

On one hand, academically developed FFs are freely available, however they are typically trained with data for just hundreds of model compounds, severely limiting the quality on vast drug-like chemical space. On the other hand, commercial FFs, such as OPLS from Schrödinger, are trained to cover a much larger chemical space, but they are only available to users with commercial licenses. In addition, the commercial FFs cannot be typically utilized in combination with academically developed FFs for biomolecules and cannot be employed in widely used simulation packages like GROMACS,[17] OpenMM,[18] AMBER,[19] CHARMM,[20] etc, limiting the development of simulation tools for the broader community. Systematic FF development to accurately cover larger chemical space is a challenging endeavor. It is a

major undertaking that requires significant scientific expertise and manual efforts in training set selection, in addition to access to high performance compute resources. It normally takes many years to develop FFs anew or to upgrade an existing one to a new version. Due to the critical role of the FF in molecular simulations, we believe it is necessary to develop a new general FF for systematic coverage of drug-like molecules and make such FF widely available to the research community to speed up scientific innovation. Recent progress of the Open Force Field (OpenFF) initiative has led to the development of more accessible open-source tools for FF development which enable wider community collaboration.[15,21–24] While OpenFF provides a new scheme, SMIRKS Native Open Force Field (SMIRNOFF),[25] and a development platform to define and build parameters to provide high coverage of chemical space, the amount of training data set with QM energy is still relatively small.[22,26]

Here, we present a new FF, XtalPi force field (XFF), designed and developed to produce higher quality and broader coverage of drug-like molecular chemical space compared to the available FFs. XFF is trained with a large set of selected functional fragments and high quality QM energies. Moreover, the parameters of XFF are Open Access and compatible with the AMBER biomolecule FFs and can be deployed in the widely used molecular simulation packages including AMBER, GROMACS, OpenMM, CHARMM etc. The remainder of the manuscript is organized as follows. In the Methods section, we describe the workflow and protocol for the FF training, QM/MM conformer validation, and free energy perturbation (FEP) benchmark. In the Results section, we show the performance of XFF in reproducing QM energies and geometries, as well as in the accuracy of FEP predictions using several benchmark sets, and we compare it with other academic and commercial FFs. In the Discussion and Conclusion sections, the performance of XFF is summarized, while the remaining problems and future developments will be discussed.

Methods

Functional form

XFF uses the Amber functional form:

$$E = \sum_i k_{r,i}(r - r_{\mathrm{eq},i})^2 + \sum_i k_{\theta,i}(\theta - \theta_{\mathrm{eq},i})^2 + \sum_i \sum_{n=1}^{4} \frac{V_{n,i}}{2}[1 + \cos(n\phi_i - \gamma_{n,i})] + \sum_{i<j}[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}] \quad (1)$$

where $k_{r,i}$ is the bond force constant, $r_{\mathrm{eq},i}$ is the bond stretching equilibrium value, $k_{\theta,i}$ is the angle force constant, $\theta_{\mathrm{eq},i}$ is the angle bending equilibrium value, $V_{n,i}$ is the torsion force constant, $\gamma_{n,i}$ is the torsion phase factor, A_{ij} and B_{ij} are Lennard-Jones (LJ) parameters, q_i and q_j are partial charges on the atom i and j, and R_{ij} is the distance between atom i and j.

Training set selection and fragmentation

The training set for XFF was generated from the ChEMBL database.[27] All molecules from the ChEMBL database were fragmented into elementary fragments, fundamental building blocks of a molecule that generally comprise 0-2 rotatable dihedrals. The elementary fragments were then recombined into one of the two types of secondary fragments: (1) chain secondary fragments and (2) ring secondary fragments. A chain secondary fragment consists of a central chain elementary fragment that is connected with other 1-3 chain or ring elementary fragments. Whereas, a ring secondary fragment, if defined as a substructure, comprising a central ring elementary fragment that is also connected with other 1-3 chain or ring elementary fragments. The details about our fragmentation method can be found in the Supporting Information and Figure S1-S3, which illustrates the procedure to generate elementary and secondary fragments. Our current fragmentation scheme does not explicitly take conjugation into account thus it may have limitations for some large conjugate

molecules. Recently, a new fragmentation method was developed that utilizes the Wiberg bond order to detect the minimum fragment needed to reproduce the chemical environment of the full molecule,[28] which can to some extent handle conjugation. After the fragment recombination process was completed, a scaffold network graph method[29] was used to cluster all fragments. Finally, we used these clusters to select nearly 55,000 elementary and secondary fragments as the training set for XFF, aiming to achieve an adequate balance between computational cost and chemical space diversity.

Torsion scan

A torsion scan workflow was developed to generate training targets. QM calculations were carried out for each model compound in the training set along the scanned torsions to obtain the energies, Hessian, and structures of the molecules. These properties were then used in the intra-molecular parameter fitting. The workflow comprised the following steps. First, the RDKit package[30] was used to convert the SMILES representation of a molecule to a 3D structure. Except for a subset of selected model compounds, protonation state and tautomer enumerations were assigned to the widely used form. Subsequently, a quick MM torsion scan using GAFF was carried out along every flexible torsion angle of the molecule. The global minima of every torsion angle were combined to produce a "global minimum" conformation of the molecule. Finally, a QM optimization of this global minimum conformation was carried out at B3LYP/6-31G* level of theory using Gaussian 16[31] and the resulting optimized geometry was used as input in the later procedures. The topology of each model compound was inferred from the QM optimized geometry and no other resonance structures were considered. Each flexible torsion was classified into three distinct categories based on the type of scan carried out around that torsion: (1) 1D, (2) pseudo 2D, or (3) 2D. Torsions in the 1D scan category were subject to independent scan around the specified dihedral over -180° to 180° with a step size of 30°. A geometry optimization of each intermediate conformation was done, constraining the specified torsion angle at the specified value, while all other degrees of

freedom (DOFs) were relaxed. Torsions in the pseudo 2D category were subjected to multiple individual 1D scans along a given torsion, each starting from different local minima along for the other flexible torsions. Finally, in the third full 2D category, constrained optimization of all combinations of the two coupled torsions was performed. A fixed step size of 30° was used for each of the three categories. For each torsion scan, first QM geometry optimization of each intermediate conformations were performed at B3LYP/6-31G* (neutral and positively charged molecules) or B3LYP/6-31+G** level of theory for negatively charged molecules using Gaussian 16,[31] followed by a single-point energy calculation at MP2/cc-pVTZ level of theory using Psi4.[32] The QM local minima of every torsion were again combined to produce "combined" conformations of the molecule, as illustrated in Figure S4. If the number of the "combined" conformations was too large, only 50 low-energy conformations were kept based on their estimated energy. The estimated energy of a "combined" conformation was calculated as the sum of the torsion strain energy over all flexible torsions. The torsion strain energy for a given dihedral angle value in a "combined" conformation was defined as the energy difference between the global minima conformation along the 1D PES for that dihedral angle and the energy of the corresponding conformation from the 1D torsion scan that has the torsion angle value corresponding to the "combined" conformation. If the QM calculation of the resulting "combined" conformation failed due to steric clash, this conformation was then discarded. These "combined" conformations underwent the same optimization and single-point calculation procedure as the torsion scan conformation except that all torsions were allowed to relax at the optimization stage. Then, for the global minimum and the "combined" conformations of each molecule, the Hessian (frequency) was computed at B3LYP/6-31+G** level to serve as target data for bond and angle parameter fitting. Finally, for each molecule, multiple "combined" conformations (<=10) were selected for QM calculation at HF/6-31G* level to generate the QM electrostatic potential for RESP[11] charge fitting.

Atom type definition

Atom type definition is at the core of FF development. Although some recent progress in FF development has eliminated the need for atom types by directly defining the chemical environment of parameter terms,[25,33,34] most of the widely used FFs for molecular modeling continue to rely on atom types to optimize the number of FF parameters required for broad chemical space coverage. To define atom types, we followed a hierarchical method that was explicitly proposed in a general FF – TEAM[35] publication, and was also actually employed in other FFs.[36] The term "hierarchical" mainly refers to the following two points: (1) during atom type definition, the environment of an atom can be introduced hierarchically; (2) all atom types can be organized in a tree structure. The atom typing scheme consists of default and extension definitions, such that the default definition only considers the properties of the atom that is being defined, while the extension definitions include additional neighboring atoms. If a more complex chemical environment is encountered and a new atom type is needed, the existing atom types can be extended to describe a wider surrounding and the subsequent new atom type can be added to the appropriate place in the tree as a new descendant of the existing type. Figure S5 of the Supporting Information provides an example showing how atom types are organized in a tree structure. Furthermore, to limit the number of parameters and prevent over-fitting, an equivalence table was also introduced to reduce an atom type to one of its ancestors in the context of certain FF term types (e.g., bond, angle, or torsion).

Bonded and non-bonded terms fitting

The bonded terms were fitted through three consecutive steps. First, the bond and angle equilibrium values r_{eq} and θ_{eq} were parametrized by gradually approaching their QM values using the method described in Wang et al.[37] to better reproduce QM optimized structures. In an ideal situation, MM optimized structures should resemble the QM structures if the corresponding QM bond and angle equilibrium values were adopted for the MM parameters.

However, in some cases, strong 1-4 electrostatic interactions can distort a MM structure, causing large deviations from the ideal QM geometries. To prevent such discrepancies, the following algorithm was introduced:

1. The initial values of $r_{i,\mathrm{eq}}$ and $\theta_{i,\mathrm{eq}}$ were taken from the QM average values over all training conformations that contain this FF term. All torsion scan and "combined" conformations were used in this stage of parameter fitting.

2. MM geometry optimizations of these conformations were carried out, and average QM/MM difference for each bond and angle term was calculated by

$$\Delta_i = \frac{1}{N_{\mathrm{conf},i}} \sum_n (x_{\mathrm{QM},i,n} - x_{\mathrm{MM},i,n}) \tag{2}$$

   where x_i is either r_i or θ_i. The summation n is over all conformations related to this term and $N_{\mathrm{conf},i}$ is the total number of such conformations.

3. The bond or angle equilibrium value $x_{\mathrm{eq},i}$ was then incremented by $x'_{\mathrm{eq},i} = x_{\mathrm{eq},i} + s\Delta_i$, where s is a scaling factor which is set to 0.2.

4. Step 2 and 3 were repeated until the following objective function was converged,

$$L_{\mathrm{ba,equil}} = w_b \sum_i \frac{\Delta_{\mathrm{bond},i}}{n_{\mathrm{bond}}} + w_a \sum_j \frac{\Delta_{\mathrm{angle},j}}{n_{\mathrm{angle}}} \tag{3}$$

   where n_{bond} and n_{angle} are the total number of bond and angle equilibrium parameters to be fitted, w_b and w_a are weight factors and were empirically set, respectively, to 3.0 and 1.0 for the fitting process.

Bond and angle force constant parameters $k_{r,i}$ and $k_{\theta,i}$ were then parametrized using a modified Seminario Hessian projection method as described in Allen et al.[38] Only the global QM minimum conformations from the torsion scan for each molecule were utilized for the

force constant fitting, and the final value for each term was averaged over all molecules that contain that term.

Finally, the torsion force constants $V_{1,i}$, $V_{2,i}$, $V_{3,i}$, and $V_{4,i}$ were parametrized using all conformations (including torsion scan and "combined" conformations) and energies from the torsion scan workflow. An objective function that calculates the difference between optimized MM energies $E_{\mathrm{MM},ij}$ and QM energies $E_{\mathrm{QM},ij}$ was introduced and was minimized by the L-BFGS-B method,[39]

$$L_{\mathrm{dihe}} = \frac{1}{N_{\mathrm{mol}}} \sum_{i}^{N_{\mathrm{mol}}} \frac{1}{N_{\mathrm{conf},i}} \sum_{j}^{N_{\mathrm{conf},i}} [(E_{\mathrm{MM},ij} - \bar{E}_{\mathrm{MM},i}) - (E_{\mathrm{QM},ij} - \bar{E}_{\mathrm{QM},i})]^2 \tag{4}$$

where the first summation is over all the molecules in the training set and N_{mol} is the total number of molecules. The second summation is over the conformations of a given molecule and $N_{\mathrm{conf},i}$ is the number of conformations for molecule i. $\bar{E}_{\mathrm{MM},i}$ and $\bar{E}_{\mathrm{QM},i}$ are the MM and QM average energy over the given molecule, respectively.

XFF directly takes LJ parameters from GAFF1.8,[1] and the electrostatic partial charges were fitted against QM electrostatic potential (ESP) using the standard RESP charge fitting workflow.[11] However, we note that XFF can easily and quickly adapt to any non-bonded parameter change since we have designed a particularly fast and efficient implementation of data structure and algorithm that can handle the intra-molecular parameter fitting workflow for an entire training set of ∼55K molecules and ∼2.8M conformations in less than two days on a machine with 53 CPU cores, which allows us to perform inter- and intra-molecular parameter fitting iteration efficiently.

Conformer geometry and energy validation

Molecule conformer energy is the most important and direct way to judge the quality of the FF parameters. To validate the conformer energy, we collected three benchmark data sets (see Table 1 for details). The first data set, referred as the Fragment Set, was derived

from the PDB[40] and Pfizer's internal compound library and was broken down into molecular fragments to reduce QM computation cost. This set was considered a challenging test set for XFF because OPLS3 energies of the molecules in this set showed large discrepancy from the corresponding QM energies. For molecules from the PDB database, an additional criterion was applied that the complex resolution is $\leq$ 2.0 Å and DPI[41] $\leq$ 0.5 Å (when the information is available). The second data set, referred as the Full-molecule Set, represents a subset of publicly available parent molecules from the Fragment Set as well as additional kinase inhibitors from the PDB. The third data set was a public data set assembled by an industrial consortium to benchmark the performance of OpenFF (OpenFF Industrial Set).[21]

Table 1: Summary of validation molecule sets

Data set Name	Origin	No. of structures	No. of conformers	Level
Fragment Set	Pfizer & PDB	2506 fragments	64,718 conformers	Fragment
Full-molecule Set	PDB	1459 molecules	20,334 conformers	Molecule
OpenFF Industrial Set	Industry	9329 molecules[a]	71448 conformers	Molecule

[a] We note that the number of molecules and conformers downloaded was different from the number reported in Ref 21, hence we listed the molecule and conformer ID we used in the benchmark on Github (see Data availability section for more details).

For the Fragment Set, the same torsion scan workflow used in the training set construction was utilized for each fragment, and the validation mainly focused on torsion profile comparison. For the Full-molecule Set, we did not carry out torsion scans due to extensive QM calculations required. Instead, we only ran QM calculations at the same level of theory as the Fragment Set on conformers generated by RDKit.[30] After QM calculations, conformers with QM energy higher than 15 kcal/mol from the minimum energy conformer of the corresponding molecule were discarded. To compare the performance of XFF with other FFs, we also carried out MM calculations with GAFF2 (version 2.11) and the latest version of OpenFF (OpenFF2.0.0, FF filename "openff_unconstrained-2.0.0.offxml", using RDKit as the backend).[26] The charge model used for XFF and GAFF2 was RESP,[11] while for OpenFF2 AM1BCC[12] charges were used to maintain consistency with the original FF development philosophy. The MM energy of the molecule was calculated by minimizing the

molecule with the corresponding FF starting from the QM optimized structures. Whenever the conformer was from a torsion scan, a dihedral restraint was also added to the molecule. Optimizations with all three FFs XFF, GAFF2, and OpenFF2 were performed with OpenMM.[18]

For the OpenFF Industrial Set, the QM, OpenFF2, and GAFF2 optimized geometries and their corresponding energies were downloaded from the QCPortal[42] using the data set name provided in Ref 21 and new calculations were only conducted for XFF using the same MM energy calculation method used for the other two data sets. We point out that the charge model used for GAFF2 in QCPortal for this data set is AM1BCC, which is different from RESP, the desirable charge model for GAFF2.[43] However, we did not re-calculate the GAFF2 results using RESP to be consistent with the results from Ref 21. One should also note that the QM level for the third data set is B3LYP-D3BJ/DZVP, which is consistent with the one used in the OpenFF training set but different from our training QM level of MP2/cc-pVTZ//B3LYP/6-31G*. This data set is a "blind" set to XFF and poses more challenges to the FF. On one hand, XFF has an advantage with the Fragment and Full-molecule Set since the QM level of these two data sets is consistent with the XFF training set. However, for the OpenFF Industrial Set, this turns to a disadvantage. Most of the molecules in this data set were not collected by us but by a few pharmaceutical companies and their focus in chemical space may be very different from that of a public database such as ChEMBL. To give the readers a general picture, Figure S7 and Table S5-S7 of the Supporting Information provides histogram distributions of the molecular weight and number of rotatable bonds for the three data sets.

The comparison was carried out to assess energetic and geometric agreement between FFs and QM. The geometry metrics included atom-wise heavy atom root mean square deviation (RMSD) and torsion fingerprint deviation (TFD)[44] between MM-optimized and QM-optimized geometry for each conformer of the molecules. The energy comparison was summarized in terms of energy correlation and energy deviation metrics. For the Fragment Set,

torsion-based Pearson correlation coefficient (R) and root mean square error (RMSE) were calculated. On the other hand, molecule-based R and conformer energy deviation ($\Delta\Delta E$) described in Ref 21 were obtained for the other two molecule-level data sets. Here, the correlation coefficient R was used instead of R^2 to avoid overestimation of the results due to negative R values. The equations to calculate the correlation and energy deviation metrics can be found in the Supporting Information.

FEP calculations

Relative binding free energy (RBFE) calculations were carried out using the internally developed XFEP platform which has already been described elsewhere,[45] hence we will only briefly detail the simulation protocol here. All simulations were performed on graphics processing unit (GPU) using the GPU-accelerated version of the AMBER20 code PMEMD.[19] The R-group substitution, heterocycle focused, charge-change data set from Lu et al.,[6] and the Merck KGaA set from Schindler et al.[46] and their corresponding initial structures from the original literature were used in the FEP benchmark study. The proteins and the ligands were modeled using AMBER ff14SB[47] and general XFF parameters respectively, and the explicit solvation environment was modeled using the TIP3P water model.[48] The ligand-protein complexes and the ligands were both solvated in an octahedral simulation box with a buffer size of 8 Å and 12 Å respectively using tLEaP. The simulation box was then neutralized by adding an appropriate number of Na^+ or Cl^- ions. The SPLIT method from Machado and Pantano[49] was used to add additional ions in cases where the ionic strength of the simulation needs to match experimental conditions. Each ligand and ligand-protein complex was initially equilibrated with the following protocol. First, a brief minimization was conducted using 100 steps of steepest descent followed by 100 steps of conjugate gradient with a Cartesian restraint of 4 kcal mol^{-1} Å^{-2} applied to the heavy atoms of the ligand and the complex. Then, the system was heated to 298.15 K in 100 ps in the NVT ensemble and was further equilibrated for 100 ps in the NPT ensemble. Finally, the restraints were

gradually removed over 400 ps again in the NPT ensemble. The final structure of the system was then extracted to construct the alchemical transformation topology with the aid of an internally developed atom mapping algorithm implemented in the XFEP platform to identify common core and softcore atoms. The hydrogen mass was increased to 3.024 amu by repartitioning the mass from the nearest-bound heavy atom.[50]

The alchemical calculations were performed in a one-stage concerted lambda scheme. During this stage, both the LJ parameters and the partial charges of the initial state were transformed into the final state with lambda windows at 0.0, 0.0479, 0.1151, 0.2063, 0.3161, 0.4374, 0.5626, 0.6839, 0.7937, 0.8849, 0.9521, and 1.0. The alchemical topology and structures prepared as described above were utilized as the initial structure for all the lambda windows. For each lambda window, we first performed minimization on the system followed by heating and pressure equilibration. Heating steps were run in the NVT ensemble, while pressure equilibration and production steps were performed in NPT with pressure maintained at 1 atm with a Monte Carlo barostat.[51] The SHAKE algorithm[52] was used to constrain the hydrogen bond length except for those bonds which connect a common core atom and a softcore atom. The non-bonded interactions were cut at 8 Å, and the particle mesh Ewald (PME) method[53] and a long-range continuum correction[54] were used for treating long-range electrostatic and LJ interactions respectively. All simulations were run using a Langevin integrator with a friction coefficient of 2 ps^{-1}, a timestep of 2 fs was used for heat and pressure equilibration while a value of 4 fs was used for production.

Following the pressure equilibration step, the final structures of all lambda windows were used to perform a Hamiltonian replica exchange molecular dynamics (HREMD)[55] production simulation, allowing all lambda windows to exchange with their neighbors. The REST2 algorithm[56] was applied to enhance the sampling process. In addition, we adopted an in-house procedure that implemented an enhanced sampling method similar to the recently published alchemical enhanced sampling (ACES)[57] to mitigate the impact of initial conformation dependence of the transforming chemical group. Specifically, the "gti_add_sc=3"

option was used to scale both the LJ and electrostatic interactions along with the alchemical lambda variable within the softcore region. All rotatable torsions within the softcore region and connecting common core and softcore region were also scaled with the lambda variable. This scheme should be equivalent to the "gti_add_sc=6" option in the ACES method, which is now readily available in AMBER22.[58] The entire HREMD production phase was run for 25 ns. Finally, the free energies and the statistical uncertainties were evaluated with the Bennett acceptance ratio (BAR) method using pymbar.[59,60] For those simulations involving a net charge change, a post-correction method was applied.[61] Recently, we have developed a new method to correct $\Delta\Delta G$ for hysteresis and convert $\Delta\Delta G$ to ΔG,[62] and this method is implemented in our XFEP platform. However, for the sake of comparison with previous studies,[6] the cycle closure correction method described in Wang et al.[63] was used throughout this work. The corrected $\Delta\Delta G$ values were then converted to ΔG by the maximum likelihood estimation method using all experimental data.[64] Three performance metrics, root mean square error (RMSE), Pearson correlation coefficient (R^2), and Kendall's rank correlation (τ), were used to compared FEP results with the experimental data. 95% confidence intervals were esimated for each metric using 1000 bootstrap samples and are reported as $x^{x_{\text{high}}}_{x_{\text{low}}}$, where x is the value from maximum likelihood estimation using the whole data set, while x_{low} and x_{high} are the value of the variable at 2.5% and 97.5% of 1000 sorted samples, respectively. To perform a fair comparison and avoid differences in details of algorithm implementation, the cycle closure corrected $\Delta\Delta G$, converted ΔG, and all performance metrics for FEP+ were re-calculated using our analysis scripts. Figure S6 provides a flowchart for a normal FEP simulation on the XFEP platform. The initial structures and results of all FEP simulations and our analysis scripts can be found on Github (see Data availability section for details).

Results

Conformer geometry and energy validation

Here, we present the results of comparison with QM for XFF, OpenFF2, and GAFF2. The results for the Fragment Set, Full-molecule Set, and OpenFF Industrial Set are shown in Figures 1, 2, and 3 respectively. On the geometry comparison tests, XFF outperforms the other FFs, showing the highest agreement with QM followed by OpenFF2 and GAFF2 on all three data sets. The percentages of conformers in the Fragment Set that have RMSD less than 0.1 Å were 61.3% for XFF, higher than the 50.6% of OpenFF2 and 40.1% of GAFF2. This number is lower for the other two data sets, which is reasonable considering RMSD is molecule-size dependent and the Full-molecule set and OpenFF Industrial set contain larger molecules. XFF predicts 19.5% and 16.3% of conformers within 0.1 Å from QM-optimized structures for the two data sets vs. 11.6% and 8.7%, respectively for OpenFF2, and 9.0% and 5.9%, respectively for GAFF2. The relative performance of different FFs based on the TFD metric shows a similar trend for all three data sets. XFF outperforms OpenFF2 and GAFF2, generating 87.6%, 76.7%, and 76.0% of the conformers with TFD < 0.05 on the three data sets. Regarding the energy metrics, XFF has a large advantage in terms of torsion-based R and RMSE on the fragment set, demonstrating the ability to correctly reproduce QM torsion profiles. Correctly predicting QM potential energy surfaces (PES) is critical in biomolecular simulations as it decides the orientation of important functional groups of molecules and thus contributes to the penalization of the host-guest interactions in the form of conformation strain energy. In Figure S8 of the Supporting Information, we also show three example torsion PESs that have a correlation coefficient R equal to 0.9, 0.8, and 0.6 respectively. The examples indicate that the R=0.9 torsion has a good overall QM and MM PES shape agreement while the other two torsions do not. The ranking of FFs is XFF $>$ OpenFF2 $>$ GAFF2 for this data set. XFF has 77.8% of torsions having a correlation higher than 0.9 with QM and 65.0% of torsions with RMSE less than 1.0 kcal/mol, while these two

numbers are only 60.9%/34.5% for OpenFF2 and 53.8%/32.4% for GAFF2. Concerning the other two molecule-level data sets, XFF still shows excellent performance on the QM/MM conformer energy correlation, while the energy deviations are closer for all three FFs. For the Full-molecule set, XFF is the best among the three, containing 43.2% of molecules having QM/MM conformer correlation larger than 0.9 and 44.3% of conformers with $|\Delta\Delta E|$ less than 1 kcal/mol. GAFF2 has a slightly higher percentage of conformers with $\Delta\Delta E$ within the range $[-1.0, 0.0)$ kcal/mol than XFF, but the total percentage with $|\Delta\Delta E| <$ 1.0 kcal/mol is 42.7%, slightly smaller than XFF. The performance trend is the same for the OpenFF Industrial set. The $R > 0.9$ percentage is 50.2% for XFF and is better than the other two FFs with a large margin. The numbers for $|\Delta\Delta E| < 1.0$ kcal/mol are close for three FFs, with XFF still the best at 50.2%, followed by 49.6% of OpenFF2 and 46.1% of GAFF2. The $\Delta\Delta E$ distribution for XFF is slightly shifted right compared with OpenFF2 and GAFF2, probably due to the fact that the QM level of the OpenFF Industrial Set is different from the other two sets as we have discussed earlier. We noticed that there are still some molecules with large $|\Delta\Delta E|$ and low Pearson R values for XFF. They will be further investigated and will be the future direction for FF refinement. In Figure S10(A) of the Supporting Information, we list some typical outlier molecules from the OpenFF Industrial Set that have large QM/MM energy discrepancy and low QM/MM energy correlation.

FEP benchmark using R-group substitution data set

The R-group substitution data set consists of eight test cases from Wang et al.:[65] BACE, CDK2, Jnk1, MCL1, p38, PTP1B, Thrombin, and Tyk2. Similar to our previous work,[45] we compared the RMSE, Pearson correlation coefficient (R^2), and Kendall's rank correlation (τ) between the cycle-closure corrected FEP $\Delta\Delta G$ results and the experimental values from XFF with the latest version of FEP+.[6] In addition, $\Delta\Delta G$ values were converted to ΔG using all experimental data and were also compared with FEP+. What differs from the previous work is that we used the general version of XFF throughout this work instead of a

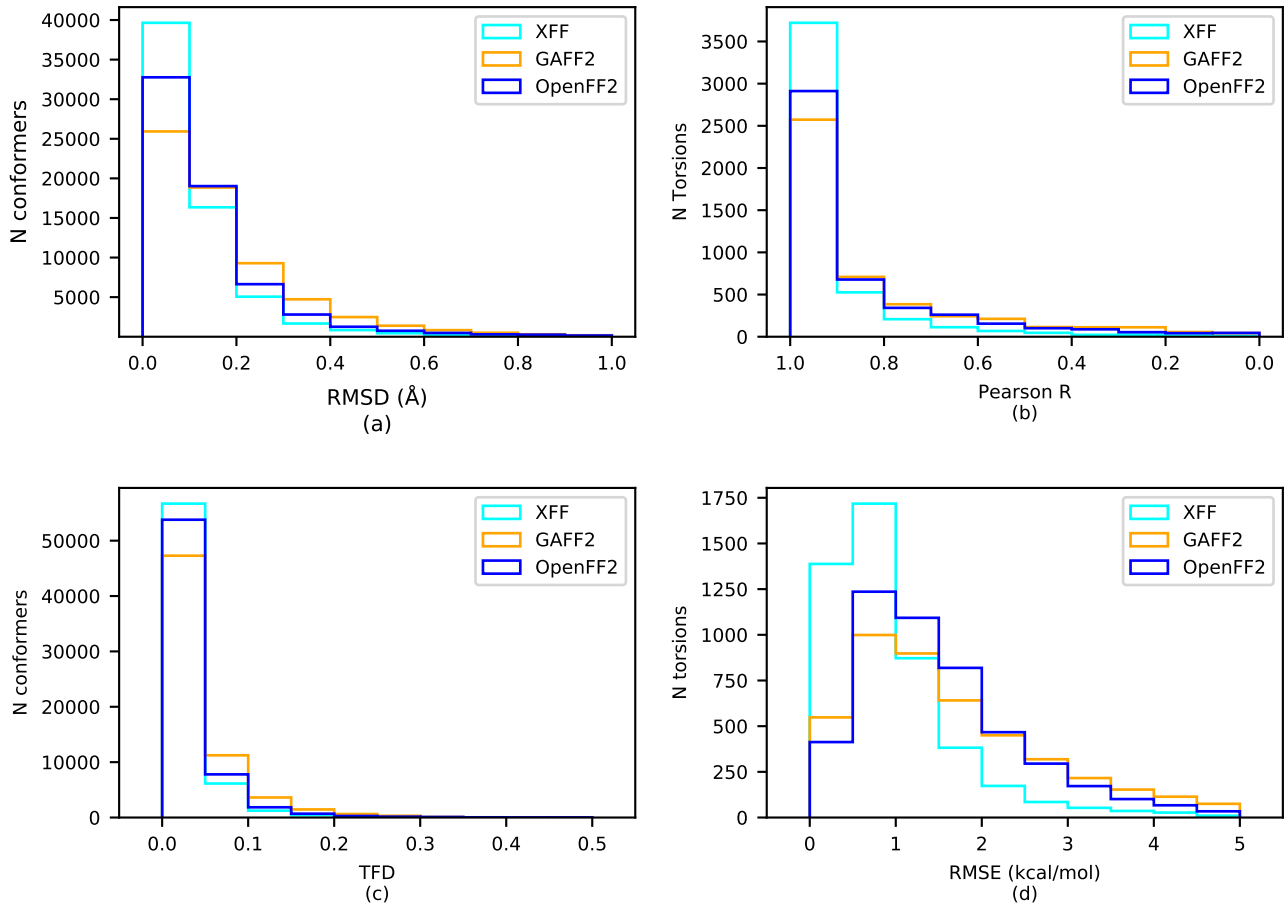

Figure 1: Histogram of validation results for XFF, GAFF2 and OpenFF2 FFs on the Fragment Set. (a) RMSD distribution between MM-optimized structures compared to QM structures. (b) Pearson correlation coefficient distribution of QM/MM potential energy surface. (c) TFD distribution between MM-optimized structures compared to QM structures. (d) RMSE distribution of QM/MM potential energy surface. Numerical data corresponding to the plot can be found in Table S8, S9, S10, and S11 of the Supporting Information.

system-specific one which includes performing refinement on each test case. Such refinement is usually not affordable for massive FEP simulations since expensive QM calculations are required. Meanwhile, the general version of the FF is fast to obtain and reflects the true quality of the FF. We performed FEP simulations for the same 333 pairs with the same direction as was done by Wang et al.,[65] and the results are shown in Table 2 and Figure 4. Overall, XFF together with our XFEP platform has a comparable performance with the latest version of FEP+ (OPLS4) on these eight test cases. The overall RMSE of XFF for

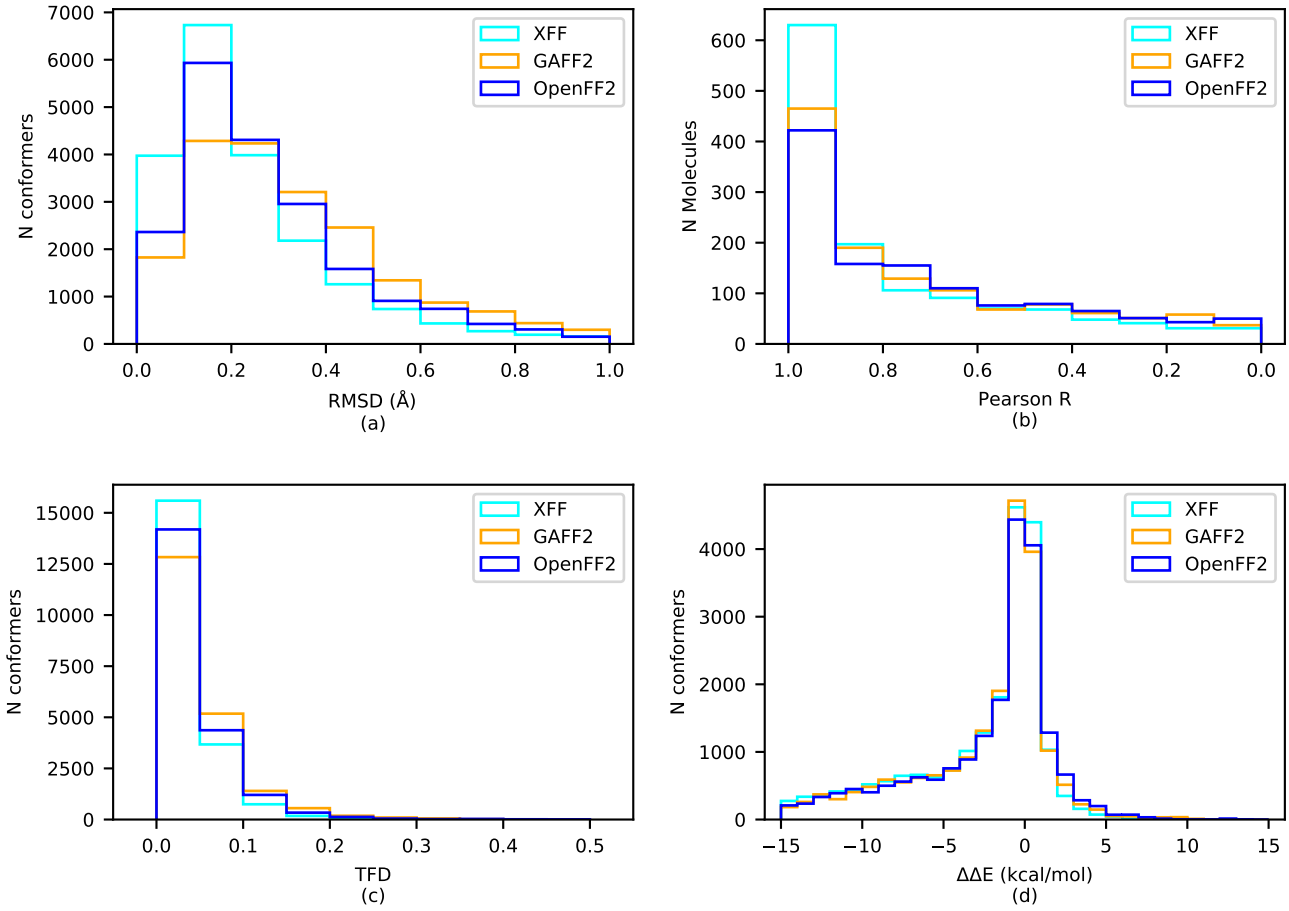

Figure 2: Histogram of validation results for XFF, GAFF2 and OpenFF2 FFs on the Full-molecule Set. (a) RMSD distribution between MM-optimized structures compared to QM structures. (b) Molecule-based Pearson correlation coefficient distribution of QM/MM energies. (c) TFD distribution between MM-optimized structures compared to QM structures. (d) Conformer-based $\Delta\Delta E$ distribution of QM/MM energies. Numerical data corresponding to the plot can be found in Table S8, S9, S10, and S12 of the Supporting Information.

$\Delta\Delta G$ is $0.94^{1.22}_{1.01}$ kcal/mol, which is slightly lower than $0.97^{1.16}_{1.00}$ kcal/mol of OPLS4. On the other hand, XFF also shows a slightly higher R^2 and τ value of $0.49^{0.49}_{0.35}$ and $0.50^{0.50}_{0.37}$ compared with $0.45^{0.47}_{0.32}$ and $0.48^{0.50}_{0.38}$ of OPLS4. After converting $\Delta\Delta G$ values to ΔG, the total RMSE of XFF is $0.81^{1.03}_{0.80}$ kcal/mol, higher than $0.76^{0.92}_{0.73}$ kcal/mol of OPLS4. However, two correlation metrics R^2 and τ for XFF were on the same level as OPLS4. The system-specific investigation shows that XFF has a much larger ΔG RMSE value of $1.22^{1.59}_{1.07}$ kcal/mol and $0.68^{1.15}_{0.57}$ kcal/mol for MCL1 and PTP1B, 0.38 kcal/mol and 0.18 kcal/mol higher than

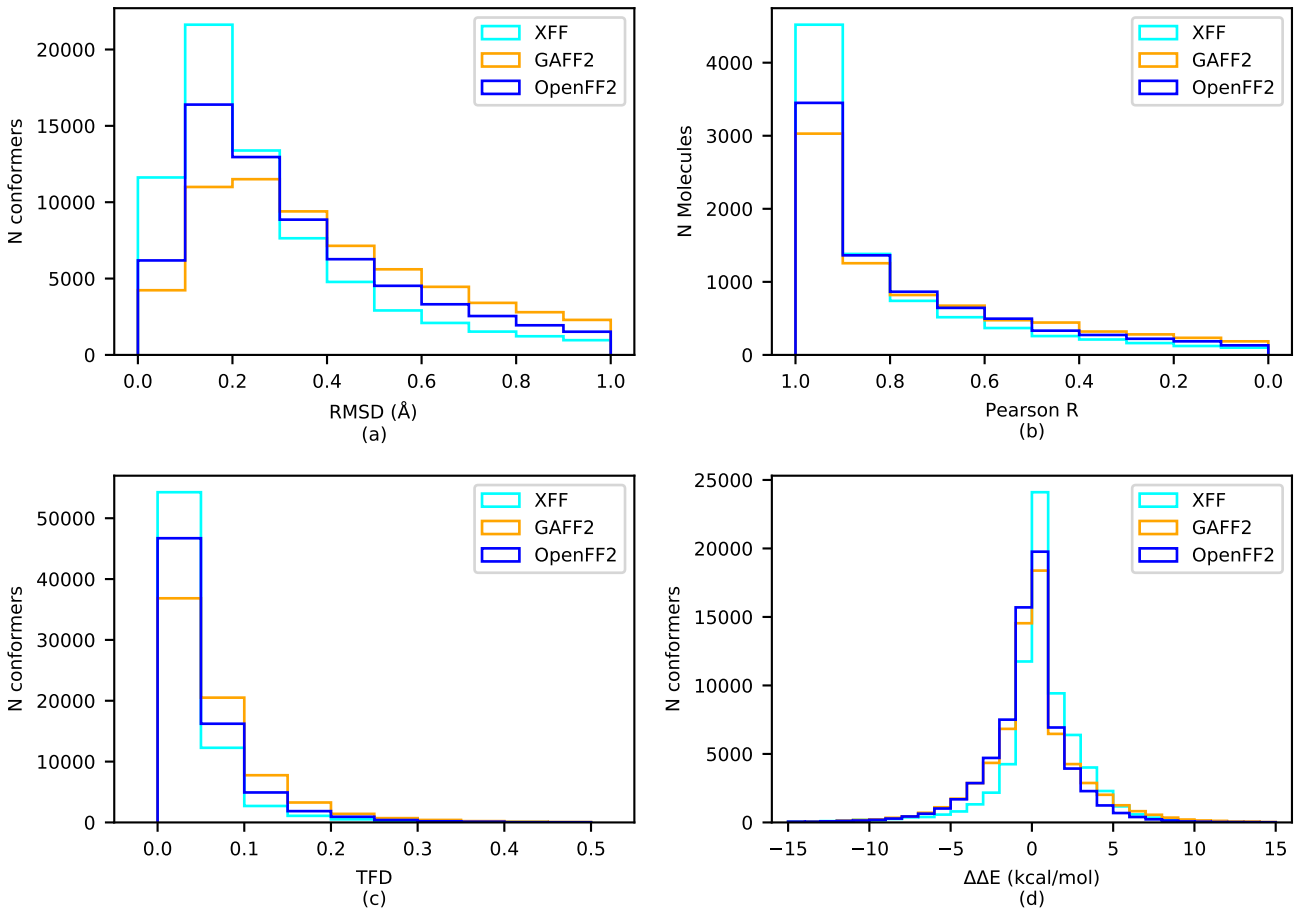

Figure 3: Histogram of validation results for XFF, GAFF2 and OpenFF2 FFs on the OpenFF Industrial Set. (a) RMSD distribution between MM-optimized structures compared to QM structures. (b) Molecule-based Pearson correlation coefficient distribution of QM/MM energies. (c) TFD distribution between MM-optimized structures compared to QM structures. (d) Conformer-based $\Delta\Delta E$ distribution of QM/MM energies. Numerical data corresponding to the plot can be found in Table S8, S9, S10, and S12 of the Supporting Information.

OPLS4, which explains XFF having an overall larger RMSE on ΔG than OPLS4. Some typical outlier transformations from MCL1 and PTP1B test cases can be found in Figure S13 of the Supporting Information.

Table 2: Summary of the performance of XFF and OPLS4 on the R-group substitution data set (RMSE in unit of kcal/mol). $\Delta\Delta G$ values are all corrected by cycle-closure correction.[63] Total RMSE values were calculated by combining all data points, while total values for R^2 and τ were calculated as the weighted average of all test cases by their corresponding number of pairs/ligands to avoid metric overestimation due to different experimental ranges.

		XFF $\Delta\Delta G$			OPLS4 $\Delta\Delta G^a$		
Targets	N_{pair}	RMSE	R^2	τ	RMSE	R^2	τ
BACE	58	$0.96^{1.22}_{0.82}$	$0.29^{0.45}_{0.07}$	$0.33^{0.47}_{0.12}$	$1.01^{1.26}_{0.86}$	$0.36^{0.52}_{0.15}$	$0.36^{0.51}_{0.20}$
CDK2	25	$0.97^{1.36}_{0.63}$	$0.58^{0.79}_{0.29}$	$0.59^{0.76}_{0.35}$	$1.12^{1.36}_{0.90}$	$0.21^{0.46}_{0.01}$	$0.31^{0.53}_{0.06}$
Jnk1	34	$0.77^{1.08}_{0.65}$	$0.39^{0.60}_{0.10}$	$0.43^{0.60}_{0.18}$	$0.89^{1.15}_{0.75}$	$0.33^{0.60}_{0.06}$	$0.40^{0.58}_{0.13}$
MCL1	71	$1.32^{1.78}_{1.27}$	$0.31^{0.40}_{0.10}$	$0.39^{0.48}_{0.20}$	$1.17^{1.43}_{1.03}$	$0.23^{0.40}_{0.07}$	$0.34^{0.46}_{0.17}$
p38	56	$0.74^{1.26}_{0.83}$	$0.71^{0.72}_{0.36}$	$0.65^{0.65}_{0.39}$	$0.84^{1.17}_{0.81}$	$0.66^{0.72}_{0.40}$	$0.64^{0.68}_{0.44}$
PTP1B	49	$0.83^{1.37}_{0.88}$	$0.73^{0.73}_{0.33}$	$0.66^{0.67}_{0.37}$	$0.71^{1.40}_{0.89}$	$0.80^{0.76}_{0.40}$	$0.74^{0.71}_{0.46}$
Thrombin	16	$0.64^{0.98}_{0.50}$	$0.13^{0.52}_{0.00}$	$0.21^{0.54}_{-0.28}$	$1.02^{1.32}_{0.73}$	$0.31^{0.69}_{0.03}$	$0.43^{0.67}_{0.01}$
Tyk2	24	$0.48^{0.60}_{0.39}$	$0.87^{0.93}_{0.73}$	$0.68^{0.81}_{0.49}$	$0.85^{1.12}_{0.61}$	$0.65^{0.82}_{0.40}$	$0.66^{0.79}_{0.43}$
Total	333	$0.94^{1.22}_{1.01}$	$0.49^{0.49}_{0.35}$	$0.50^{0.50}_{0.37}$	$0.97^{1.16}_{1.00}$	$0.45^{0.47}_{0.32}$	$0.48^{0.50}_{0.38}$
		XFF ΔG			OPLS4 ΔG^a		
Targets	N_{lig}	RMSE	R^2	τ	RMSE	R^2	τ
BACE	36	$0.87^{1.10}_{0.74}$	$0.59^{0.74}_{0.29}$	$0.56^{0.69}_{0.34}$	$0.87^{1.10}_{0.68}$	$0.51^{0.72}_{0.23}$	$0.46^{0.65}_{0.24}$
CDK2	16	$0.76^{1.13}_{0.46}$	$0.61^{0.88}_{0.26}$	$0.53^{0.75}_{0.22}$	$0.91^{1.16}_{0.61}$	$0.41^{0.75}_{0.09}$	$0.43^{0.69}_{0.05}$
Jnk1	21	$0.53^{0.75}_{0.42}$	$0.76^{0.87}_{0.55}$	$0.69^{0.82}_{0.49}$	$0.73^{0.95}_{0.55}$	$0.66^{0.85}_{0.38}$	$0.68^{0.80}_{0.43}$
MCL1	42	$1.22^{1.59}_{1.07}$	$0.38^{0.54}_{0.14}$	$0.44^{0.57}_{0.21}$	$0.84^{1.14}_{0.68}$	$0.56^{0.69}_{0.35}$	$0.58^{0.68}_{0.39}$
p38	34	$0.59^{1.04}_{0.58}$	$0.68^{0.73}_{0.28}$	$0.69^{0.70}_{0.42}$	$0.78^{1.11}_{0.68}$	$0.49^{0.65}_{0.18}$	$0.55^{0.65}_{0.28}$
PTP1B	23	$0.68^{1.15}_{0.57}$	$0.73^{0.81}_{0.21}$	$0.65^{0.74}_{0.33}$	$0.50^{0.96}_{0.46}$	$0.85^{0.89}_{0.50}$	$0.73^{0.81}_{0.46}$
Thrombin	11	$0.42^{0.65}_{0.29}$	$0.41^{0.76}_{0.01}$	$0.38^{0.71}_{-0.15}$	$0.57^{0.78}_{0.32}$	$0.67^{0.92}_{0.14}$	$0.60^{0.86}_{0.14}$
Tyk2	16	$0.36^{0.46}_{0.27}$	$0.95^{0.98}_{0.89}$	$0.82^{0.92}_{0.56}$	$0.45^{0.68}_{0.26}$	$0.89^{0.96}_{0.75}$	$0.82^{0.92}_{0.55}$
Total	199	$0.81^{1.03}_{0.80}$	$0.62^{0.64}_{0.46}$	$0.59^{0.61}_{0.45}$	$0.76^{0.92}_{0.73}$	$0.63^{0.69}_{0.51}$	$0.61^{0.64}_{0.47}$

[a] ref 6.

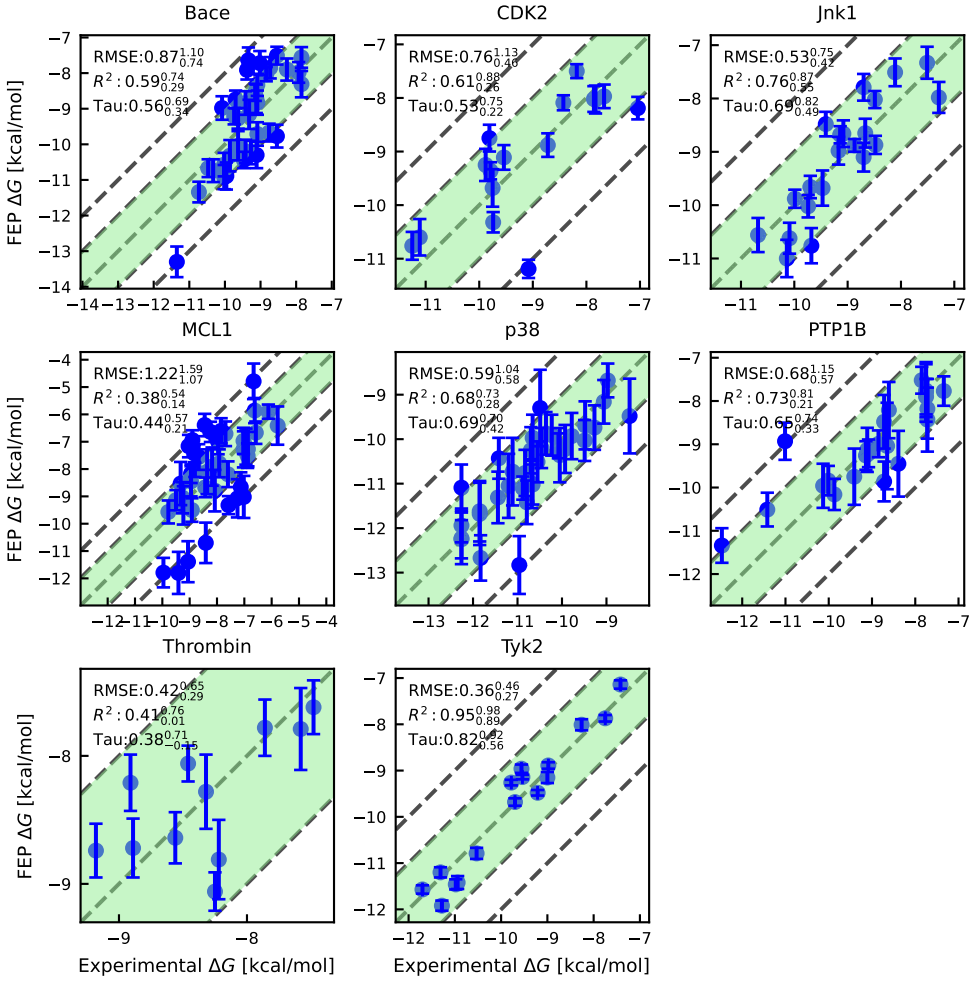

Figure 4: Correlation between FEP predicted binding free energy (ΔG) results and experimental values for the eight test cases from Wang et al.[65] The green shaded area delimited by the dashed lines encloses all data points that have FEP results within 1 kcal/mol from the corresponding experimental values. The outer dashed lines correspond to $y = x + 2$ and $y = x - 2$ respectively.

FEP benchmark using heterocycle focused data set

Four public test cases, CHK1, FXA, BACE core, and BACE heterocycles, from Roos et al.[66] were considered for further FEP validation. This data set is more challenging than the R-group substitution set for a general FF as it is mainly composed of heterocyclic cores with

diverse heteroatoms and ring substitutions at different positions, and thus requires a more accurate description of torsion profiles and electrostatic potentials. Since the number of perturbation pairs in each subset of these 4 test cases is relatively small and it is inappropriate to combine data from different experimental conditions and calculate correlation metrics, we only show the RMSE for $\Delta\Delta G$ pairs here. The performance of XFF on this data set is shown in Table 3. Overall, XFF has a comparable performance with OPLS4 except for the BACE core system where XFF has a very poor performance. One possible explanation for the poor performance is the charge model since the BACE core test case comprises molecules with unusual amidine cores, which pose a great challenge for the ESP description. Currently, the charge model used in XFF is a standard RESP fitting protocol[11] starting from a DFT-optimized structure. The RESP charge model is well known for some deficiencies such as conformation dependence and may produce non-physical charges. Thus, it may be necessary to remove the bias of the RESP charges by introducing multiple conformations and adding additional restraints.[66,67] Another possible source of error is the protonation states of the catalytic ASP diad. As pointed out by Roos et al.,[66] the electron-rich substitution on the bicyclic compound may shift the protonation state of ASP228 to neutral. In Ref 66, the authors concluded that with the presence of compound **23**, which has a OMe substitution, ASP228 was in neutral form. However, there are other compounds in the series which also have a high eletronegativity substitution, such as fluorine for compounds **19** and **20**, chlorine for compound **21**, OH for compound **22**, and CF_3 for compound **24** (shown in Figure 5). Thus, we calculated the pK_a shift of ASP228 with all bicyclic compounds in the BACE core system using the same method as in Ref 66. The pK_a values of ASP228 with bicyclic compounds can be found in the Table S1 of the Supporting Information. Based on our calculations, the pK_a values of ASP228 together with compounds **19**, **20**, **21**, **22**, **23**, and **24** came closer to or larger than the experimental pH of 4.5, which means ASP228 should be neutral or at least partially neutral. Hence, we applied the protein residue pK_a correction to the $\Delta\Delta G$ values as described in Ref 66. As expected, we did observe improvements for

some pairs. For example, the $\Delta\Delta G$ between **21** and **23** was decreased from 1.39 kcal/mol to -0.12 kcal/mol, which is in line with the experimental measurement of -0.1 kcal/mol. The $\Delta\Delta G$ between **22** and **23** was improved by more than 2 kcal/mol, from 2.36 kcal/mol to -0.05 kcal/mol, which is closer to the experimental value of 0.1 kcal/mol. However, we found no improvements or even deterioration for several other pairs, and the overall performance of the BACE core system did not improve (see Table S2 from the Supporting Information). The main reason performance didn't improve significantly is that pK_{a} corrections for perturbations involving compound **24** went in the opposite direction and severely impaired the accuracy. This may be attributed to the fact that some ligands (especially **24**) may not be stable under the current double-deprotonated structure and further investigation is needed to clarify the issue.

Table 3: Summary of the performance of XFF and OPLS4 on heterocycle focused data set (RMSE in unit of kcal/mol). $\Delta\Delta G$ values are all corrected by cycle-closure correction method. Total RMSE was calculated by combining pair data from all test cases.

		XFF $\Delta\Delta G$	OPLS4 $\Delta\Delta G^a$
Targets	N_{lig}	RMSE	RMSE
CHK1	20	$1.43^{2.17}_{0.62}$	$0.96^{1.26}_{0.66}$
FXA	41	$1.28^{1.61}_{1.14}$	$1.35^{1.63}_{1.13}$
BACE core	23	$2.05^{2.83}_{1.82}$	$1.21^{1.58}_{0.99}$
BACE heterocycles	21	$0.94^{1.22}_{0.70}$	$1.20^{1.59}_{0.85}$
Total	105	$1.45^{1.81}_{1.35}$	$1.24^{1.44}_{1.12}$

[a] ref 6.

FEP benchmark using charge-change data set

Eleven test cases obtained from Lu et al.,[6] which involve net charge change between perturbation pairs, were used to validate XFF. Since most of the perturbed pairs in this set have a net charge change, a post-correction method[61] was applied to correct for the finite size effect. For some of the ligands that have a potential protonation center, the pK_{a} effect must also be taken into account during FEP simulations. Consequently, the pK_{a} values of

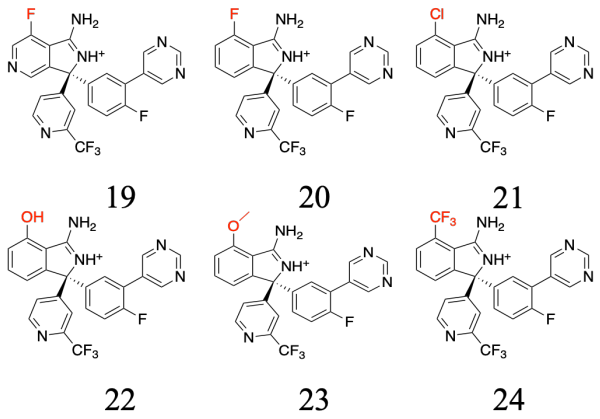

Figure 5: Compounds for BACE core system with an electron rich substitution (shown in red) on the bicyclic ring.

these ligands were calculated by Marvin[68] and are shown in Table S3 of the Supporting Information. If one ligand has a non-negligible protonation state, either the deprotonated or protonated form of the ligand was first used in the perturbation pair calculation depending on the pairs designed in the previous work,[6] then the RBFE between the deprotonated and protonated forms of the ligand was also calculated. Afterwards, the cycle closure correction was applied, and the final RBFE was calculated by applying the pK_a correction method described in Chen et al.[69] We noticed that the authors of Ref 6 did not provide their calculated pK_a values, thus it is unclear how to reproduce their RMSE values. To have a fair comparison, we recalculated the RMSE value for each test case for OPLS4 using Marvin predicted pK_a and the raw RBFE values provided in Ref 6. When we were calculating the performance metrics such as RMSE, only pK_a corrected $\Delta\Delta G$ values were compared with the experimental ones. As mentioned by the authors in Ref 69, the ionic strength may affect the results for some test cases, hence we also tested adding additional salts to match the ionic strength of the experimental condition. The results are shown in Table 4. Since the number of ligands in each test case is also relatively small, only $\Delta\Delta G$ RMSE is shown here.

For the CDK2 test case, we noticed that the carboxylic group of ligand **39** is close to residue ASP87 and may shift the pK_a value of ASP87, even neutralizing it. Thus, we also applied the protein residue pK_a correction mentioned previously to the pairs involving ligand

Table 4: Summary of the performance of XFF and OPLS4 on charge change data set (RMSE in unit of kcal/mol). The column with label "salt" results from simulations with additional salts to match the experimental buffer concentration.

			XFF $\Delta\Delta G$	OPLS4 $\Delta\Delta G^a$	XFF $\Delta\Delta G$ (salt)
Targets	N_{lig}	Exp. pH^b	RMSE	RMSE	RMSE
CDK2	3	7.5	$0.97^{1.36}_{0.23}$	$0.67^{1.10}_{0.33}$	$1.22^{1.75}_{0.55}$
DLK	5	7.5	$1.04^{2.21}_{0.44}$	$0.66^{2.82}_{0.38}$	$1.22^{3.91}_{0.80}$
EGFR	5	7.5	$1.52^{2.45}_{0.95}$	$1.53^{2.34}_{0.98}$	$1.44^{3.90}_{1.36}$
EPHX2	4	7.0	$0.57^{2.16}_{0.46}$	$1.36^{1.90}_{0.63}$	$1.19^{2.75}_{0.75}$
IRAK4	9	7.2	$1.94^{3.40}_{1.57}$	$1.28^{1.90}_{0.94}$	$2.07^{3.78}_{1.82}$
ITK	4	7.2	$2.00^{3.13}_{1.09}$	$1.11^{1.59}_{0.71}$	$1.72^{2.42}_{0.93}$
JAK1	6	7.2	$1.09^{1.83}_{0.63}$	$1.23^{2.70}_{0.75}$	$1.09^{1.82}_{0.66}$
JNK1	3	7.2	$0.78^{3.06}_{0.36}$	$0.70^{2.07}_{0.27}$	$0.34^{4.78}_{0.48}$
PTP1B	3	7.0	$1.13^{1.58}_{0.41}$	$1.49^{2.56}_{0.60}$	$0.89^{2.09}_{0.38}$
Thrombin	6	7.4	$1.58^{2.81}_{1.10}$	$1.40^{1.86}_{0.87}$	$1.53^{2.18}_{0.98}$
TYK2	5	7.0	$0.74^{1.84}_{0.45}$	$1.10^{1.73}_{0.52}$	$0.34^{0.65}_{0.12}$
Total	55		$1.45^{2.19}_{1.45}$	$1.24^{1.68}_{1.17}$	$1.46^{2.51}_{1.63}$

[a] ref 6. [b] ref 69.

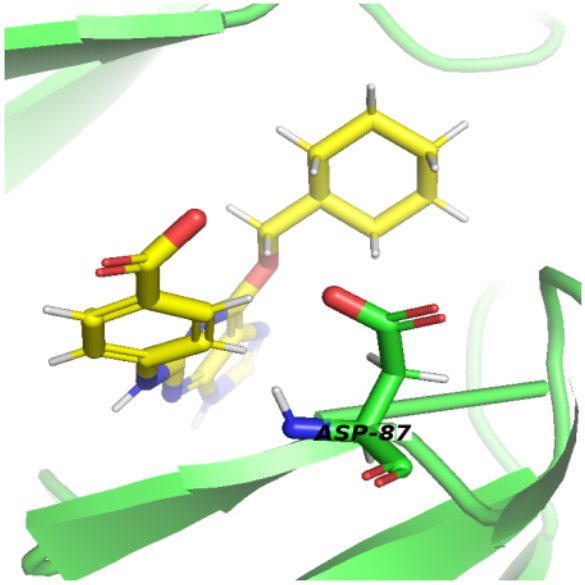

Figure 6: The carboxylic group of ligand **39** in CDK2 test case is in proximity of ASP87, which may shift the pK_{a} of the protein residue ASP87.

39. The uncorrected RBFE between the deprotonated form of **39** and **25** we calculated is −3.12 kcal/mol, which is more than 1.5 kcal/mol deviation from the experimental value

of −1.5 kcal/mol. After the pK_a correction to the ligand, the $\Delta\Delta G$ value was improved to −2.88 kcal/mol. The $\Delta\Delta G$ was further improved to −2.56 kcal/mol with the protein residue pK_a correction. Similarly, the original relative binding free energy between **39** and **34** is −3.32 kcal/mol and the experimental value is −1.5 kcal/mol. The $\Delta\Delta G$ value became −2.77 kcal/mol after the ligand and residue pK_a correction was applied. In both cases, the two pK_a corrections only brought about 0.3 kcal/mol improvement each to the $\Delta\Delta G$ value, which does not constitute a significant difference considering the typical error of FEP simulations, demonstrating that deprotonated forms are dominant in both ligand **39** and ASP87.

In some other cases, the pK_a correction did improve FEP predictions. For example, in the JAK1 test case, the pK_a of compounds **18** and **16** was predicted to be 8.4 and 10.4 respectively. Under the experimental pH conditions (pH=7.2), they should be both in the protonated state. The RBFE between **18** charged and **16** charged from FEP is −0.20 kcal/mol, about 1 kcal/mol deviation from the experimental value of 0.8 kcal/mol. Also, the predicted $\Delta\Delta G$ between **18** charged and **jmc_34** is −3.02 kcal/mol while its corresponding experimental measurement is −1.8 kcal/mol. The pK_a correction was applied on both compounds **18** and **16**, after which the $\Delta\Delta G$ between **18** and **16** was calculated to be 0.60 kcal/mol, leaving only about 0.2 kcal/mol discrepancy from the experiment. The $\Delta\Delta G$ between **18** and **jmc_34** was also improved by 0.8 kcal/mol to −2.23 kcal/mol. A detailed analysis shows that the predicted $\Delta\Delta G$ between **18** charged and **18** neutral is −2.28 kcal/mol, which means that the pK_a of compound **18** was shifted by about 1.7 pK_a units to 6.7 upon binding thus both forms should be considered. On the other hand, the predicted $\Delta\Delta G$ between **16** charged and **16** neutral is only −0.60 kcal/mol and the pK_a of **16** was only shifted by 0.4, which has a negligible effect on the RBFE. In another case, compound **19a** from the JNK1 system has a pK_a value of 8.5. The RBFE between **6a** and the charged form of **19a** was 2.04 kcal/mol and the experimental value is 1.6 kcal/mol. The calculated $\Delta\Delta G$ between the neutral and charged form of **19a** was 1.76 kcal/mol, decreasing

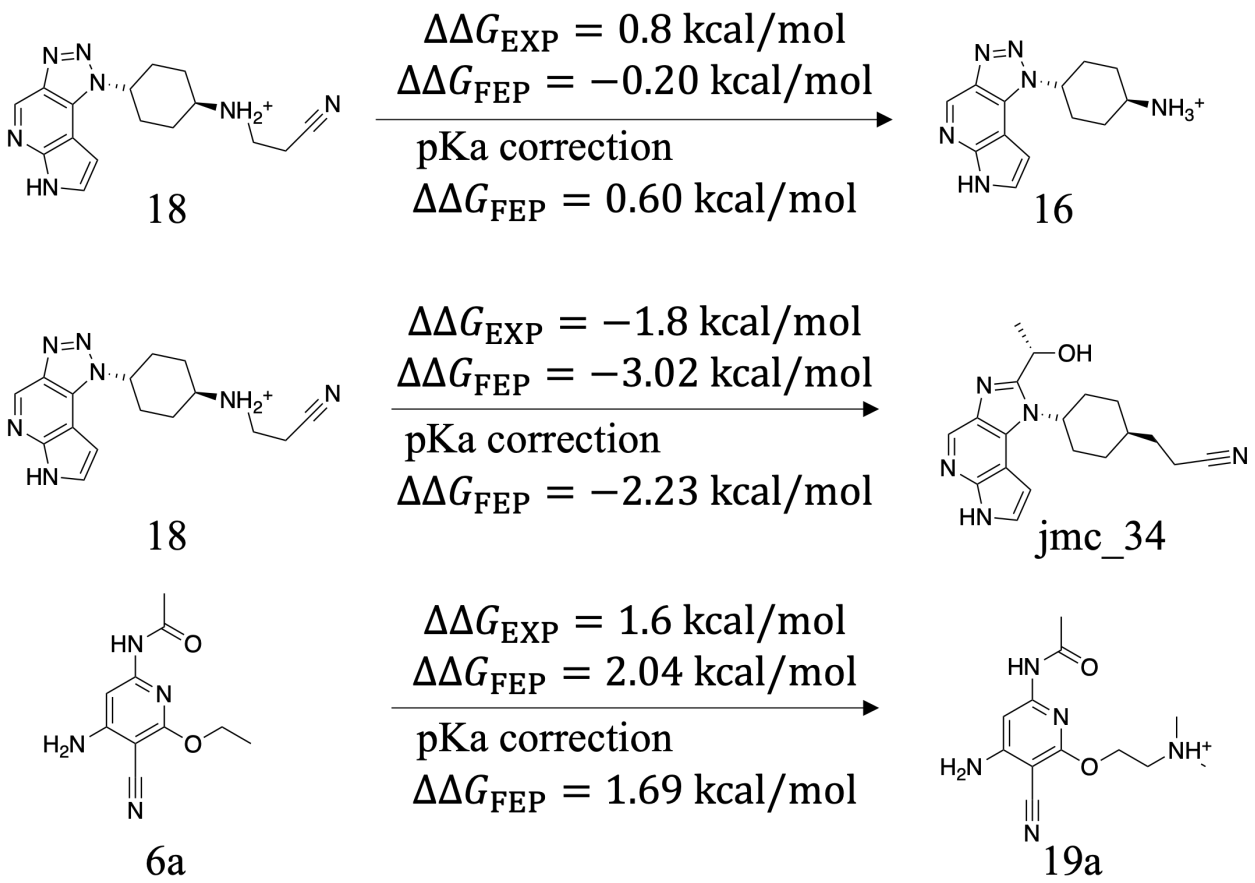

Figure 7: For perturbation pair 18 to 16 and 18 to jmc_34 in JAK1 system and pair 6a to 19a in JNK1 system, the pK_a correction improves their relative binding free energy predictions.

the pK_a of **19a** to 7.2. After the pK_a correction, the $\Delta\Delta G$ between **6a** and **19a** was predicted to be 1.69 kcal/mol, only less than 0.1 kcal/mol higher than the experimental value.

As pointed out in Ref 69, for some test cases, adding additional salt ions to the simulation can improve FEP predictions. Ideally, adding additional ions representing the experimental condition can stabilize the structures compared to only neutralizing the system with counter ions. As a result, we also tested adding Na^+ and Cl^- ions to match the experimental ionic strength, and the results are shown in Table 4. Our simulations indicated that the ITK, JNK1, PTP1B and TYK2 test cases significantly benefited from additional salts. The RMSE for $\Delta\Delta G$ predictions for these test cases was improved by ∼0.3 kcal/mol. However, in many other cases, additional salts deteriorated the simulation accuracy. For example, in

EPHX2, the $\Delta\Delta G$ RMSE is increased by more than 0.6 kcal/mol. In CDK2 and DLK test cases, the $\Delta\Delta G$ RMSE have a $\sim$0.2 kcal/mol increase. For other test systems, we did not observe significant change. Overall, the impact of adding additional ions to the simulations is difficult to predict, and it should be evaluated on a case-by-case basis.

FEP benchmark using Merck KGaA data set

This data set was assembled by Merck KGaA [46] and contains 8 test cases representative of real drug discovery projects. The protein and ligand input structures were directly taken from Schindler et al. [46] except for PFKFB3 and TNKS2, and the simulations were performed on the same 540 perturbation pairs. To be consistent with the benchmark procedure adopted in the previous sections, the production phase of the simulation was still set to 25 ns and the results were compared with the 20 ns simulation from Schindler et al. For the EG5 test case, the protein structure with remodeled loop was used. For the PFKFB3 test case, we found that the structure from Schindler et al. was not stable due to the highly charged environment near the substrate and ATP binding pockets. Indeed, the natural substrate fructose-6-phosphate (F6P) and ATP molecules both have negative charges which can help neutralize and stabilize the positive charges in the pocket. Since protein structure preparation is highly important for FEP accuracy and we want to separate its effects from the FF quality, we tried to determine the best protein input structure for the PFKFB3 test case. Based on the description of the experimental assay [70] and other crystal structures of this system, we decided to add a substrate molecule F6P and a phosphate anion in the FEP input structure. The coordinates of the phosphate anion were modeled using the coordinates of the pyrophosphate anion in the crystal structure 6HVI, while the coordinates of the F6P molecule were obtained by aligning crystal structure 2DWP to 6HVI. To validate our modification, we ran 100 ns MD simulations for the protein-ligand complex with and without the two co-factors. The complex stability was improved by the presence of the two co-factors (see Figure S11(C) and (D) of the Supporting Information). Detailed analysis and FEP results using the original structure

from Schindler et al. for PFKFB3 can be found in Figure S11 of the Supporting Information.

For the TNKS2 test case, two series of major outliers were observed if the original structure from Schindler et al. was used. One of the two series was associated with adding a fluorine or chlorine atom to the 6-position of 2-arylquinazolin-4-ones core. Our FEP simulations underestimated the potency of compounds containing fluorine or chlorine substitutions, meaning such substitutions were predicted to be not tolerated, which is in contradiction with the experimental data. After careful examination, we noticed that a sulfate anion is present near the binding pocket in the original structure (PDB code: 4UI5) used by Schindler et al. We searched the PDB database for other TNKS2 structures and found many other structures with a sulfate anion at the almost identical position. The sulfate anion has an interaction with a LYS residue and may also have an indirect impact on a GLU residue that forms a salt bridge with LYS. According to the crystallographic conditions,[71] the sulfate anion probably comes from Li_2SO_4 which serves as a precipitant to facilitate the crystallization process.[72,73] The concentration of Li_2SO_4 added to the solution is 0.2 M, which is beyond the physiological condition. It is clearly an artifact of crystallization and no reagent containing sulfate ion was present in the experimental assay[74] of the TNKS2 test case. Thus, we highly suspect that the crystal structure 4UI5 does not accurately represent the real structure in the experimental assay conditions. As a result, we searched TNKS2 crystal structures that do not contain sulfate anion and obtained two candidate structures (PDB code: 4PNN and 3MHK). They both have similar structures for the LYS and GLU salt bridge and are different from 4UI5. However, these two structures still have a large structural difference for residues from PHE1044 to MET1054 (conforming to the residue numbering in 3MHK), which may have an impact on FEP accuracy. In the end, 3MHK was chosen as structure for FEP calculations due to it having less interactions with other units in the crystal lattice packing in the region named above which, in turn, points to the fact that it may better represent the ligand-bound structure in solution. After introducing the new structure, the potency changes by adding a fluorine or chlorine atom were mostly predicted accurately by FEP. A comparison of 4UI5

and 3MHK shows that the GLU residue moves slightly aside in 3MHK, which makes room for the accommodation of a fluorine or chlorine atom (shown in Figure 9). The other series of outliers is related to transformations involving a net charge change, which was in agreement with what Schindler et al. found. We hypothesized that those charged molecules are actually in the neutral form upon binding to the protein. Consequently, we conducted the same pK_a correction method used in the charge-change data set and the results improved. The calculations also show that the pK_a of the charged ligands are indeed shifted to 5.1-7.8 in the protein-ligand complex (see Table S4 in the Supporting Information for more details), which means the neutral form cannot be ignored considering the experimental pH of 7.7. Comparison of results using the original structure and method from Schindler et al. and the new ones in this work can be found in Figure S12 of the Supporting Information.

The overall results of the Merck KGaA set are shown in Table 5 and Figure 8. Overall, XFF achieved comparable performance with the results from Schindler et al. For $\Delta\Delta G$ predictions, XFF has a total RMSE value of $1.22^{1.55}_{1.36}$ kcal/mol, Pearson correlation R^2 of $0.50^{0.47}_{0.35}$, and ranking correlation τ of $0.52^{0.50}_{0.41}$, which is very close to the total results from FEP+ ($1.24^{1.57}_{1.37}$, $0.44^{0.42}_{0.30}$, $0.45^{0.45}_{0.35}$). After converting $\Delta\Delta G$ to ΔG, the RMSE, R^2, and τ for XFF are, respectively, $0.99^{1.19}_{0.98}$ kcal/mol, $0.50^{0.55}_{0.36}$, and $0.51^{0.54}_{0.40}$, which is also on the same level as FEP+ ($1.06^{1.27}_{1.04}$, $0.46^{0.51}_{0.33}$, $0.49^{0.51}_{0.38}$). A system-specific analysis shows that XFF has a relatively larger error for the HIF-2α and c-Met test cases. For HIF-2α, the major outliers involve a ring-closure transformation that coverts a cyano and methyl group to a five-member ring. Analysis of the c-Met test case shows that most FEP outliers were related to molecules having a thiolactone structure, which is not very common and its FF parameters may not be well trained in XFF. Some examples of typical outlier perturbations can be found in Figure S13 of the Supporting Information.

Table 5: Summary of the performance of XFF and FEP+ on the Merck KGaA data set (RMSE in unit of kcal/mol). All performance metrics were obtained using the same method as described in previous sections.

		XFF $\Delta\Delta G$			FEP+ $\Delta\Delta G^a$		
Targets	N_{pair}	RMSE	R^2	τ	RMSE	R^2	τ
CDK8	54	$1.46^{1.80}_{1.24}$	$0.37^{0.52}_{0.18}$	$0.48^{0.60}_{0.31}$	$1.50^{1.92}_{1.35}$	$0.37^{0.50}_{0.14}$	$0.44^{0.57}_{0.25}$
Eg5	65	$0.96^{1.55}_{1.09}$	$0.58^{0.59}_{0.25}$	$0.55^{0.58}_{0.32}$	$1.17^{1.77}_{1.17}$	$0.36^{0.42}_{0.13}$	$0.46^{0.49}_{0.24}$
HIF-2α	80	$1.64^{2.20}_{1.65}$	$0.48^{0.55}_{0.25}$	$0.52^{0.56}_{0.36}$	$1.34^{2.09}_{1.42}$	$0.41^{0.45}_{0.11}$	$0.47^{0.49}_{0.23}$
c-Met	57	$1.37^{2.12}_{1.36}$	$0.55^{0.64}_{0.23}$	$0.56^{0.61}_{0.32}$	$1.16^{1.80}_{1.26}$	$0.67^{0.68}_{0.38}$	$0.62^{0.65}_{0.41}$
SYK	101	$1.06^{1.46}_{1.07}$	$0.33^{0.44}_{0.11}$	$0.37^{0.44}_{0.17}$	$1.25^{1.58}_{1.21}$	$0.20^{0.35}_{0.03}$	$0.19^{0.33}_{0.02}$
TNKS2[b]	60	$0.95^{1.43}_{0.92}$	$0.58^{0.66}_{0.27}$	$0.58^{0.62}_{0.35}$	$1.01^{1.28}_{0.87}$	$0.53^{0.67}_{0.30}$	$0.49^{0.61}_{0.32}$
PFKFB3[b]	67	$1.11^{1.33}_{1.00}$	$0.61^{0.71}_{0.44}$	$0.60^{0.67}_{0.48}$	$1.21^{1.50}_{1.02}$	$0.69^{0.78}_{0.54}$	$0.65^{0.73}_{0.54}$
SHP-2	56	$0.97^{1.61}_{1.11}$	$0.59^{0.58}_{0.22}$	$0.59^{0.59}_{0.34}$	$1.19^{1.88}_{1.30}$	$0.45^{0.51}_{0.12}$	$0.43^{0.51}_{0.19}$
Total	540	$1.22^{1.55}_{1.36}$	$0.50^{0.47}_{0.35}$	$0.52^{0.50}_{0.41}$	$1.24^{1.57}_{1.37}$	$0.44^{0.42}_{0.30}$	$0.45^{0.45}_{0.35}$
		XFF ΔG			FEP+ ΔG^a		
Targets	N_{lig}	RMSE	R^2	τ	RMSE	R^2	τ
CDK8	33	$1.09^{1.44}_{0.79}$	$0.44^{0.68}_{0.19}$	$0.50^{0.65}_{0.29}$	$1.25^{1.67}_{0.94}$	$0.38^{0.65}_{0.11}$	$0.52^{0.66}_{0.27}$
Eg5	28	$0.67^{1.04}_{0.61}$	$0.62^{0.77}_{0.15}$	$0.51^{0.63}_{0.17}$	$0.80^{1.19}_{0.62}$	$0.54^{0.71}_{0.23}$	$0.57^{0.68}_{0.30}$
HIF-2α	42	$1.33^{1.81}_{1.13}$	$0.34^{0.52}_{0.11}$	$0.44^{0.55}_{0.23}$	$1.05^{1.64}_{0.86}$	$0.33^{0.57}_{0.01}$	$0.41^{0.54}_{0.11}$
c-Met	24	$1.06^{1.55}_{0.84}$	$0.70^{0.84}_{0.38}$	$0.67^{0.78}_{0.39}$	$0.90^{1.27}_{0.77}$	$0.83^{0.90}_{0.62}$	$0.75^{0.84}_{0.58}$
SYK	44	$1.05^{1.36}_{0.88}$	$0.31^{0.53}_{0.06}$	$0.35^{0.51}_{0.14}$	$1.11^{1.40}_{0.93}$	$0.26^{0.52}_{0.02}$	$0.27^{0.45}_{0.05}$
TNKS2[b]	27	$0.88^{1.30}_{0.64}$	$0.52^{0.78}_{0.13}$	$0.59^{0.74}_{0.29}$	$1.31^{1.60}_{1.02}$	$0.24^{0.53}_{0.01}$	$0.32^{0.56}_{0.01}$
PFKFB3[b]	40	$0.81^{1.01}_{0.68}$	$0.59^{0.73}_{0.39}$	$0.56^{0.69}_{0.41}$	$1.05^{1.24}_{0.89}$	$0.67^{0.80}_{0.50}$	$0.64^{0.73}_{0.50}$
SHP-2	26	$0.65^{1.03}_{0.64}$	$0.69^{0.75}_{0.21}$	$0.59^{0.71}_{0.24}$	$0.86^{1.33}_{0.76}$	$0.55^{0.69}_{0.19}$	$0.57^{0.68}_{0.27}$
Total	264	$0.99^{1.19}_{0.98}$	$0.50^{0.55}_{0.36}$	$0.51^{0.54}_{0.40}$	$1.06^{1.27}_{1.04}$	$0.46^{0.51}_{0.33}$	$0.49^{0.51}_{0.38}$

[a] ref 46. [b] XFF results using the new structures and method from this work.

Discussion

In this study, we have completed a total number of 1079 RBFE transformations, and the general accuracy of the pairwise affinity prediction was calculated. In addition, $\Delta\Delta G$ values were converted to ΔG for the R-group substitution and Merck KGaA data sets, and their overall ΔG RMSE was also calculated. The cumulative RMSE estimate was obtained using results from the calculations without additional ions for the charge-change data set and results using the new structures and method introduced in this work for the PFKFB3 and TNKS2 test cases in the Merck KGaA data set. As can be seen from Table 6, the overall

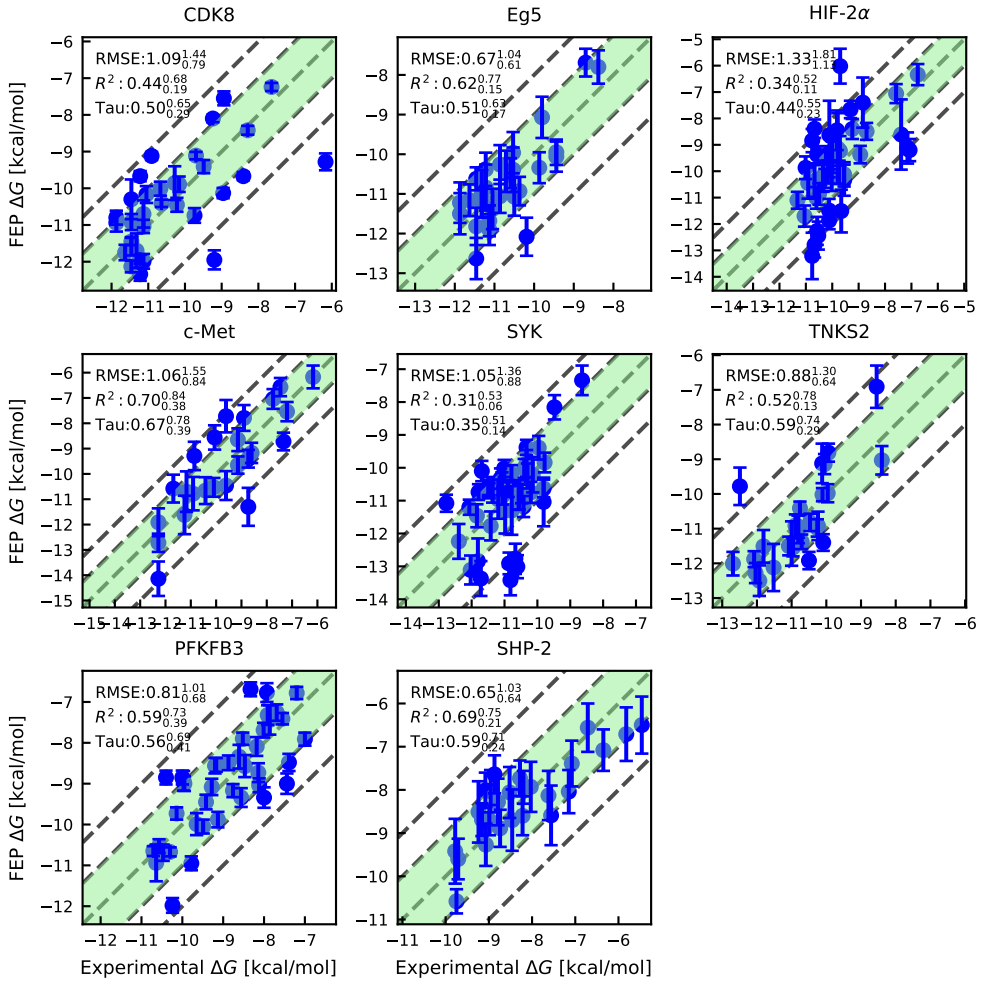

Figure 8: Correlation between FEP predicted binding free energy (ΔG) results and experimental values for the 8 test cases of the Merck KGaA data set. The green shaded area delimited by the dashed lines encloses all data points that have a FEP result within 1 kcal/mol from the corresponding experimental value. The outer dashed lines correspond to $y = x + 2$ and $y = x - 2$ respectively.

RMSE of 1079 pairs for XFF is $1.19^{1.48}_{1.33}$ kcal/mol, which is comparable with $1.16^{1.40}_{1.27}$ kcal/mol for FEP+. The accuracy for 463 ΔG values is $0.92^{1.10}_{0.94}$ kcal/mol and is comparable to the value of $0.94^{1.11}_{0.94}$ kcal/mol reported for FEP+. One should also note the compared FEP+ results for the Merck KGaA data set used the OPLS3e FF while OPLS4 was used for the other data sets. As a consequence, the above results may not accurately represent the current

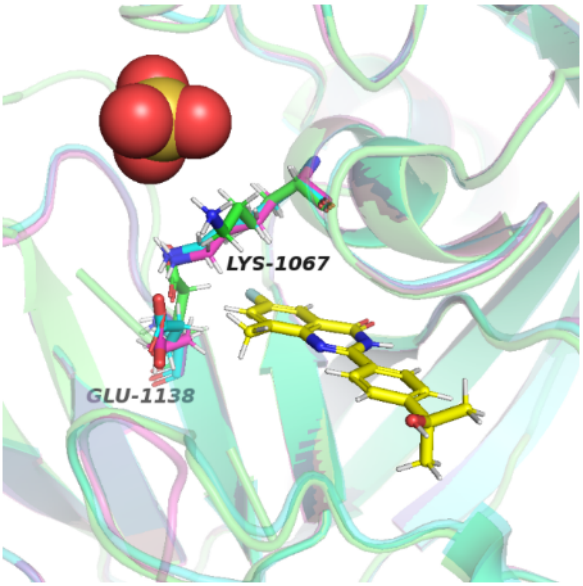

Figure 9: Comparison of protein structures 4UI5 (green), 3MHK (cyan), and 4PNN (magenta). The sulfate anion exists in 4UI5 but is absent in 3MHK and 4PNN. The LYS residue interacts with the sulfate anion in 4UI5. In 3MHK and 4PNN, due to lack of such interaction, the LYS residue is pointing to another direction. The position of the GLU residue in 3MHK and 4PNN is also different from 4UI5. Ligand **5m** is shown in yellow to illustrate that the fluorine substitution is no longer impeded by the GLU residue.

quality of the latest version of the commercial series of OPLS FFs. In Figure 10, we also show the histogram distribution of $\Delta\Delta G$ prediction error. In comparison to FEP+, XFF has 13 more pairs with prediction error smaller than 0.5 kcal/mol, but also 14 more pairs with error larger than 2.5 kcal/mol. In terms of ΔG error, XFF has 11 more molecules with prediction error smaller than 0.5 kcal/mol, and 1 less molecule with an error larger than 2.5 kcal/mol. Large outliers undermined the general performance of XFF and will be the focus of future improvements. Recently, Hahn et al.[75] investigated the performance of several academic FFs on 22 protein-ligand binding affinity targets, 16 of which overlap with the ones considered in this work. This allows us to compare XFF with various academic FFs (see Figure S14 of the Supporting Information). In general, XFF has better performance than the public FFs in most of the 16 data sets. In the next paragraphs, we will briefly discuss some aspects that can potentially enhance the quality of the XFF.

The vast "drug-like" chemical space is believed to be of the order of 10^{60} molecules.[76,77]

Table 6: Total $\Delta\Delta G$ and ΔG RMSE in this study. $\Delta\Delta G$ values are corrected with the cycle-closure correction method. (RMSE unit: kcal/mol)

N_{pair}	$\Delta\Delta G$ RMSE	
	XFF	FEP+
1079	$1.19^{1.48}_{1.33}$	$1.16^{1.40}_{1.27}$
N_{lig}	ΔG RMSE	
	XFF	FEP+
463	$0.92^{1.10}_{0.94}$	$0.94^{1.11}_{0.94}$

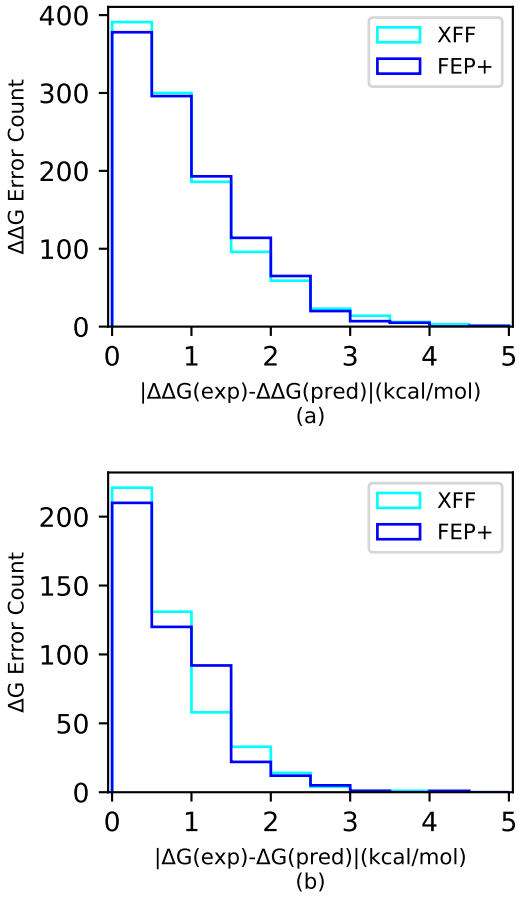

Figure 10: Histogram of the error of (a) $\Delta\Delta G$ and (b) ΔG for XFF and FEP+. Numerical values corresponding to this figure can be found in Table S13 and S14 of the Supporting Information.

The torsion term is often called a "residue-collector" term since it must remedy the inaccurate description of the electrostatic and steric interactions of the other terms. Thus, torsional

terms are often quite sensitive to a change in the chemical environment and their transferability can be questionable. One way to solve this problem would be to define more elaborate torsion types. For example, in the current version of the commercial OPLS FF, the number of torsion parameters has been expanded to ∼150K.[66] In Table 7, we also listed the number of bonded parameters for XFF and compared it with OPLS3e. However, expanding numbers of atom types and FF parameters has several drawbacks. First, a more complicated torsion type may not generalize very well and it may have disastrous behavior when the chemical environment is slightly changed. Second, as one defines more atom types, the number of torsional parameters increases combinatorially, requiring in turn a huge amount of training data. The availability of computational resources often limits the size of the training set and this inevitably leads to inadequate characterization of some of the torsional parameters. An alternative approach is bespoke fitting, meaning re-parametrization of the FF parameters for every molecule each time an FEP simulation is carried out. Many useful tools have been developed to aid the parametrization process, such as FFBuilder,[78] OpenFF BespokeFit,[22] and QUBEKit,[79] to name a few. However, bespoke fitting is demanding in terms of computational resources if a full QM level is used. To alleviate this problem some recent studies have investigated the use of less costly (but potentially less accurate) methods such as QM machine learning potential[80] or semi-empirical QM;[22] however the improvement to FEP accuracy has not yet been widely evaluated. In addition to the general version of XFF that is used throughout this work, we also developed an in-house bespoke fitting workflow based on our cloud-computing platform to refine FF quality on specific sets of molecules. The torsion scan workflow we use for the bespoke fitting is the same as the one used for the FF development. To reduce the total waiting time, we designed a system that can schedule enough cloud resources so that the QM calculations on all torsional conformations of a molecule can be performed in parallel (see Figure S9 for more details). Usually, the fitting of molecules for one batch of FEP calculations (usually tens or a few hundred molecules) can be done in 12-24 hours. Although the time required is longer than the typical FEP simulation time on

our platform (3-4 hours), it might still be worthwhile to use this workflow in some important instances such as confirming top molecules from FEP calculations using the general version of FF before they are sent for synthesis. In Figure S10(B) of the Supporting Information, we show how the QM/MM energy deviation and correlation are improved for typical outlier molecules in the OpenFF Industrial Set after applying our bespoke fitting workflow. Additionally, in relation to torsion parameters, we note that the number of torsion parameters in XFF related to a proper rotatable bond is only less than half of the total number of torsion parameters. The rest are parameters for non-rotatable bonds, for example, those torsions whose center bond is a double bond or in a ring. The force constant values of these parameters barely change in our current parametrization scheme and their values are very close to the initial values, which come from GAFF1.8. We are currently investigating a new method to parametrize non-rotatable torsions, especially those involving a saturated ring.

Table 7: Comparison of number of bonded parameters for XFF and OPLS3e. The number in parathesis denotes the number of torsion parameters related to a rotatable bond (single bond, not in a ring).

parameter type	XFF	OPLS3e[22,66]
bond	4,102	1,187
angle	13,892	15,235
torsion	96,098	146,669
	(46,774)	

As we have discussed previously, the RESP charge fitting scheme has some fundamental deficiencies. The ESP-based charges are well known for their conformation dependency and for being ill-defined for buried atoms, all of which is detrimental to the accuracy of molecular simulations, especially RBFE calculations. Although the penalty function used in the RESP[11] method alleviates these two problems to some extent, we still found that in some cases the charges of molecules heavily depend on the initial pose and there may be poor transferability of charges for the same functional group between a pair of molecules. These phenomena introduce extra noise into FEP calculations and are more severe for some polar groups such as charged amine and sulfonamide. Two examples of issues related to RESP

charges identified in our internal projects can be found in Figure S15 of the Supporting Information. Roos et al.,[66] tried to deal with these problems by using constrained optimization on a bio-active conformation of a molecule and adding a bond charge correction as an additional layer of restraint to the charge fitting procedure. Recently, Janeček et al.[67] have studied using a Hessian matrix of the objective function with respect to the charges to automatically scale up the restraint weight of the ill-defined charges to a predefined reference charge. Other approaches include using 10 conformers having electrostatically least-interacting functional groups to reduce the conformation dependence of ESP-based charges.[81] Currently, we are also investigating new charge models that are more robust with respect to the input ligand pose and more transferable for polar functional groups.

The point charge model may also not be suited to properly describe charge anisotropy. For example, there is a depletion of electron density on the back side of a halogen bond (Cl, Br, and I), which is often called a "sigma-hole".[82] It has positive electrostatic potential and may have a favorable interaction with atoms bearing a lone pair. However, the point charge model only has a negative partial charge on the halogen atom and it cannot capture the attractive behavior between the halogen atom and an electron-donating atom. Many FFs have implemented an off-site positive point charge along the elongated line of the C-X bond and showed improvements in free energy calculations.[78,83–85] Our force field XFF does not currently include any off-site charges but their inclusion will also be the subject of future investigations.

Optimization of LJ parameters is another potential direction for future improvements. Currently, the LJ parameters of XFF are directly taken from GAFF1.8.[1] Traditionally, the LJ parameters are trained against QM[86] and/or experimental pure liquid density and heat of vaporization.[10] Recently, introducing condensed-phase mixture properties to optimize LJ parameters was proposed.[87] Other advanced methods try to use atoms-in-molecule (AIM) electron density partitioning to derive parameters.[88] The progress in software tools now allows, at least in principle, to adjust LJ parameters[15,89] or define atom types[90] in an automatic

fashion. All these developments have paved the way for better descriptions of non-bonded interactions of molecules.

Finally, it has become more and more popular to evaluate the quality of small molecule FFs by comparing FEP predicted affinities with experimentally measured ones. However, FF quality is not the only factor contributing to FEP accuracy and we believe that it is important for FF developers to rule out the impact from non-FF contributions when one uses the accuracy of FEP predictions to guide the direction of FF improvements. In this work, we have demonstrated that considering protein residue and ligand protonation state changes, and using protein structures that better reflect experimental conditions can both improve FEP accuracy. In the PFKFB3 and TNKS2 test cases, using the new protein structures to set up simulations significantly improved the FEP results. We point out that the performance of other FFs may also improve using our new structures, and this renders the current comparison not entirely fair. Due to limited time and resources, we only investigated structure preparation for these two cases, but potential refinements may also be possible for other cases. Selecting an appropriate protein structure for FEP simulations is far from trivial and may require input from biochemists, structural biologists, and computational chemists. It still remains a challenge to define appropriate best practices for protein structure selection and refinement and we think the whole community may work collaboratively towards improving this fundamental aspect of FEP applications.

Conclusion

In this work, we developed an AMBER-consistent small molecule FF, XFF, whose parameters are readily available and Open Access to everyone in the field. The extensive validations on QM/MM conformer comparison and FEP calculations demonstrated the wide coverage of chemical space and the accuracy of XFF. The results were encouraging and show that XFF provides an alternative to the current academic and commercial FFs for molecular simulations

in drug discovery projects. In particular, XFF has the advantages of covering a broad range of chemical space and having its parameters freely available. Our high-throughput cloud-based bespoke fitting workflow can bring higher accuracy to FEP simulations for specific molecules. In the future, we will focus on improving the quality of the non-bonded parameters, including refinement of the RESP fitting protocol, investigating the role of off-site charge centers, and optimizing LJ parameters. We hope that our FF will prove useful for the community and will provide improvement to the quality of biomolecular simulations.

Data availability

The input structures, XFF parameters, and results for all FEP simulations, are available on Github `https://github.com/XtalPi-XFF/2023_XFF_paper` under an MIT license. The list of molecule and conformer IDs from the OpenFF Industrial set used in this work and their corresponding XFF parameters and results are also available on Github. For the Fragment Set and Full-molecule Set, data is also available for molecules in the public domain. The XFF parameters and parameter assigning tool can be accessed via a web server `https://xff.xtalpi.com`.

Associated content

The Supporting Information contains (1) detailed explanation of the fragmentation algorithm (2) illustrations for molecule fragmentation, "combined" conformation generation, and atom type definition; (3) the equations to calculate the correlation and energy deviation metrics for conformer energy validation; (4) additional notes on FEP calculations and a flowchart for an FEP calculation on the XFEP platform; (5) molecular weight and number of rotatable bonds distribution for the Fragment Set, Full-molecule Set and OpenFF Industrial Set; (6) Three example torsion PESs from the Fragment Set that have R equal to 0.9, 0.8, and 0.6. (7) a flowchart showing our bespoke fitting workflow; (8) Typical outlier molecules in the

conformer energy validation and their FF refinement with bespoke fitting workflow. (9) pK_a of ASP228 in complex with bicyclic compound in the BACE core set and its correction to FEP $\Delta\Delta G$; (10) ligand pK_a predicted by Marvin for compounds in the charge-change set; (11) comparison of FEP results using the original structure from Ref 46 and the new structures and method from this work for the PFKFB3 and TNKS2 test cases; (12) pK_a of charged ligands in the TNKS2 test case and pK_a shift upon binding to the protein; (13) example outlier transformations; (14) comparison of XFF with academic FFs; (15) two examples of issues with respect to the RESP charge fitting scheme; (16) numerical data for all the histogram plots. (PDF)

Acknowledgement

The authors acknowledge computing support from Dr. Zhe Shen, Dr. Jin Zhang, Ms. Guo Wang, Mr. Chunan Wu, Mr. Yongpan Chen, and Ms. Yunfei Zhou. BX, XW, JZ, DF, PZ, JM, SW and MY are partially supported by grants from Shenzhen Science and Technology Program (KQTD20210811090114013) and Trillion Compound Library Construction Enabling New Drug Discovery (XMHT20220104035). We would like to acknowledge Nicholas Labello for the computing infrastructure support.

References

(1) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(2) Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-acid-specific Protein Backbone Parameters Trained Against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2019**, *16*, 528–552.

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

(4) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D. CHARMM36m: an Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14*, 71–73.

(5) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(6) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *J. Chem. Theory Comput.* **2021**, *17*, 4291–4300.

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

(8) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **2015**, *11*, 3499–3509.

(9) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Cryst.* **2016**, *B72*, 171–179.

(10) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(11) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: the RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(12) Araz Jakalian, C. I. B., David B. Jack Fast, Efficient Generation of High-quality Atomic Charges. AM1-BCC model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(13) Li, P.; Song, L. F.; Merz, K. M. J. Parameterization of Highly Charged Metal Ions Using the 12-6-4 LJ-Type Nonbonded Model in Explicit Water. *J. Phys. Chem. B* **2015**, *119*, 883–895.

(14) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 2553–2558.

(15) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

(16) He, X.; Walker, B.; Man, V. H.; Ren, P.; Wang, J. Recent Progress in General Force Fields of Small Molecules. *Curr. Opin. Struct. Biol* **2022**, *72*, 187–193.

(17) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations Through Multi-level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(18) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

(19) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee Jr, T. D.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 5595–5623.

(20) Brooks, B. R.; III, C. L. B.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Caflisch, S. B. A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Gao, S. F. J.; Hodoscek, M.; W. Im, K. K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comp. Chem.* **2009**, *30*, 1545–1615.

(21) D'Amore, L.; Hahn, D. F.; Dotson, D. L.; Horton, J. T.; Anwar, J.; Craig, I.; Fox, T.; Gobbi, A.; Lakkaraju, S. K.; Lucas, X.; Meier, K.; Mobley, D. L.; Narayanan, A.; Schindler, C. E. M.; Swope, W. C.; in 't Veld, P. J.; Wagner, J.; Xue, B.; Tresadern, G. Collaborative Assessment of Molecular Geometries and Energies from the Open Force Field. *J. Chem. Inf. Model.* **2022**, *62*, 6094–6104.

(22) Horton, J. T.; Boothroyd, S.; Wagner, J.; Mitchell, J. A.; Gokey, T.; Dotson, D. L.; Behara, P. K.; Ramaswamy, V. K.; Mackey, M.; Chodera, J. D.; Anwar, J.; Mobley, D. L.; Cole, D. J. Open Force Field BespokeFit: Automating Bespoke Torsion Parametrization at Scale. *J. Chem. Inf. Model.* **2022**, *62*, 5622–5633.

(23) Boothroyd, S.; Wang, L.-P.; Mobley, D. L.; Chodera, J. D.; Shirts, M. R. Open Force Field Evaluator: An Automated, Efficient, and Scalable Framework for the Estimation of Physical Properties from Molecular Simulation. *J. Chem. Theory Comput.* **2022**, *18*, 3566–3576.

(24) Qiu, Y.; Smith, D. G. A.; Stern, C. D.; Feng, M.; Jang, H.; Wang, L.-P. Driving Torsion Scans with Wavefront Propagation. *J. Chem. Phys.* **2020**, *152*, 244116.

(25) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076–6092.

(26) Boothroyd, S.; Behara, P. K.; Madin, O. C.; Hahn, D. F.; Jang, H.; Gapsys, V.; Wagner, J. R.; Horton, J. T.; Dotson, D. L.; Thompson, M. W.; Maat, J.; Gokey, T.; Wang, L.-P.; Cole, D. J.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Shirts, M. R.; Mobley, D. L. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theory Comput.* **2023**, *19*, 3251–3275.

(27) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(28) Stern, C. D.; Bayly, C. I.; Smith, D. G. A.; Fass, J.; Wang, L.-P.; Mobley, D. L.; Chodera, J. D. Capturing non-local through-bond effects in molecular mechanics force fields I: Fragmenting molecules for quantum chemical torsion scans. 2022, bioRxiv; `https://www.biorxiv.org/content/early/2022/01/30/2020.08.27.270934`, (accessed September 18, 2023).

(29) Scott, O. B.; Edith Chan, A. ScaffoldGraph: an Open-source Library for the Generation and Analysis of Molecular Scaffold Networks and Scaffold Trees. *Bioinformatics* **2020**, *36*, 3930–3931.

(30) RDKit: Open-source cheminformatics. `http://www.rdkit.org`, (accessed September 13, 2023).

(31) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.;

Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision C.01. Gaussian Inc. Wallingford CT, 2016.

(32) Turney, J. M.; Simmonett, A. C.; Parrish, R. M.; Hohenstein, E. G.; Evangelista, F. A.; Fermann, J. T.; Mintz, B. J.; Burns, L. A.; Wilke, J. J.; Abrams, M. L.; Russ, N. J.; Leininger, M. L.; Janssen, C. L.; Seidl, E. T.; Allen, W. D.; Schaefer, H. F.; King, R. A.; Valeev, E. F.; Sherrill, C. D.; Crawford, T. D. Psi4: an Open-source ab initio Electronic Structure Program. *Wiley Interdiscip. Rev. Comput.* **2012**, *2*, 556–565.

(33) Wei, W.; Champion, C.; Liu, Z.; Barigye, S. J.; Labute, P.; Moitessier, N. Torsional Energy Barriers of Biaryls could be Predicted by Electron Richness/Deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-type Independence. *J Chem. Inf. Model.* **2019**, *59*, 4764–4777.

(34) Champion, C.; Barigye, S. J.; Wei, W.; Liu, Z.; Labute, P.; Moitessier, N. Atom Type Independent Modeling of the Conformational Energy of Benzylic, Allylic, and Other Bonds Adjacent to Conjugated Systems. *J Chem. Inf. Model.* **2019**, *59*, 4750–4763.

(35) Jin, Z.; Yang, C.; Cao, F.; Li, F.; Jing, Z.; Chen, L.; Shen, Z.; Xin, L.; Tong, S.; Sun, H.

Hierarchical Atom Type Definitions and Extensible All-Atom Force Fields. *J. Comput. Chem.* **2016**, *37*, 653–664.

(36) Vanommeslaeghe, K.; MacKerell Jr, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J Chem. Inf. Model.* **2012**, *52*, 3144–3154.

(37) Wang, R.; Ozhgibesov, M.; Hirao, H. Analytical Hessian Fitting Schemes for Efficient Determination of Force-constant Parameters in Molecular Mechanics. *J. Comput. Chem.* **2018**, *39*, 307–318.

(38) Allen, A. E.; Payne, M. C.; Cole, D. J. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.* **2018**, *14*, 274–281.

(39) Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208.

(40) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(41) Cruickshank, D. W. J. Remarks About Protein Structure Precision. *Acta Cryst.* **1999**, *D55*, 583–601.

(42) QCPortal. `http://docs.qcarchive.molssi.org/projects/QCPortal/en/latest/`, (accessed May 10, 2023).

(43) Orr, A. A.; Sharif, S.; Wang, J.; MacKerell, A. D. J. Preserving the Integrity of Empirical Force Fields. *J. Chem. Inf. Model.* **2022**, *62*, 3825–3831.

(44) Schulz-Gasch, T.; Schärfer, C.; Guba, W.; Rarey, M. TFD: Torsion Fingerprints As a

New Measure To Compare Small Molecule Conformations. *J. Chem. Inf. Model.* **2012**, *52*, 1499–1512.

(45) Lin, Z.; Zou, J.; Liu, S.; Peng, C.; Li, Z.; Wan, X.; Fang, D.; Yin, J.; Gobbo, G.; Chen, Y.; Ma, J.; Wen, S.; Zhang, P.; Yang, M. A Clound Computing Platform for Scalable Relative and Absolute Binding Free Energy Predictions: New Opportunities and Challenges for Drug Discovery. *J. Chem. Inf. Model.* **2021**, *61*, 2720–2732.

(46) Schindler, C. E. M.; Baumann, H.; Blum, A.; Böse, D.; Buchstaller, H.-P.; Burgdorf, L.; Cappel, D.; Chekler, E.; Czodrowski, P.; Dorsch, D.; Eguida, M. K. I.; Follows, B.; Fuchß, T.; Grädler, U.; Gunera, J.; Johnson, T.; Jorand Lebrun, C.; Karra, S.; Klein, M.; Knehans, T.; Koetzner, L.; Krier, M.; Leiendecker, M.; Leuthner, B.; Li, L.; Mochalkin, I.; Musil, D.; Neagu, C.; Rippmann, F.; Schiemann, K.; Schulz, R.; Steinbrecher, T.; Tanzer, E.-M.; Unzue Lopez, A.; Viacava Follis, A.; Wegener, A.; Kuhn, D. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *J. Chem. Inf. Model.* **2020**, *60*, 5457–5474.

(47) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(48) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(49) Machado, M. R.; Pantano, S. Split the Charge Difference in Two! A Rule of Thumb for Adding Proper Amounts of Ions in MD Simulations. *J. Chem. Theory Comput.* **2020**, *16*, 1367–1372.

(50) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecu-

lar Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864–1874.

(51) Feenstra, K. A.; Hess, B.; Berendsen, H. J. Improving Efficiency of Large Time-scale Molecular Dynamics Simulations of Hydrogen-rich Systems. *J. Comput. Chem.* **1999**, *20*, 786–798.

(52) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(53) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An Nlog (N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(54) Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. S. Accurate and Efficient Corrections for Missing Dispersion Interactions in Molecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 13052–13063.

(55) Sugita, Y.; Okamoto, Y. Replica-exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(56) Wang, L.; Friesner, R. A.; Berne, B. Replica Exchange with Solute Scaling: a More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431–9438.

(57) Lee, T.-S.; Tsai, H.-C.; Ganguly, A.; York, D. M. ACES: Optimized Alchemically Enhanced Sampling. *J. Chem. Theory Comput.* **2023**, *19*, 472–487.

(58) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; T. E. Cheatham, I.; Cisneros, G. A.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kaymak, M. C.; King, E.;

Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shajan, A.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiong, Y.; Xue, Y.; York, D. M.; Zhao, S.; Kollman, P. A. *Amber 2022*. University of California, San Francisco, 2022.

(59) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

(60) Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration. *J. Chem. Theory Comput.* **2010**, *6*, 1018–1027.

(61) Chen, W.; Deng, Y.; Russell, E.; Wu, Y.; Abel, R.; Wang, L. Accurate Calculation of Relative Binding Free Energies Between Ligands with Different Net Charges. *J. Chem. Theory Comput.* **2018**, *14*, 6346–6358.

(62) Yang, Q.; Burchett, W.; Steeno, G. S.; Liu, S.; Yang, M.; Mobley, D. L.; Hou, X. Optimal Designs for Pairwise Calculation: An Application to Free Energy Perturbation in Minimizing Prediction Variability. *J. Comput. Chem.* **2020**, *41*, 247–257.

(63) Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *J. Chem. Theory Comput.* **2013**, *9*, 1282–1293.

(64) Xu, H. Optimal Measurement Network of Pairwise Differences. *J. Chem. Inf. Model.* **2019**, *59*, 4720–4728.

(65) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(66) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1863–1874.

(67) Janeček, M.; Kührová, P.; Mlýnský, V.; Otyepka, M.; Šponer, J.; Banáš, P. W-RESP: Well-Restrained Electrostatic Potential-Derived Charges. Revisiting the Charge Derivation Model. *J. Chem. Theory Comput.* **2021**, *17*, 3495–3509.

(68) Cherinka, B.; Andrews, B. H.; Sánchez-Gallego, J.; Brownstein, J.; Argudo-Fernández, M.; Blanton, M.; Bundy, K.; Jones, A.; Masters, K.; Law, D. R.; Rowlands, K.; Weijmans, A.-M.; Westfall, K.; Yan, R. Marvin: A Tool Kit for Streamlined Access and Visualization of the SDSS-IV MaNGA Data Set. **2019**, *158*, 74.

(69) Chen, W.; Deng, Y.; Russell, E.; Wu, Y.; Abel, R.; Wang, L. Accurate Calculation of Relative Binding Free Energies between Ligands with Different Net Charges. *J. Chem. Theory Comput.* **2018**, *14*, 6346–6358.

(70) Boutard, N.; Białas, A.; Sabiniarz, A.; Guzik, P.; Banaszak, K.; Biela, A.; Bień, M.; Buda, A.; Bugaj, B.; Cieluch, E.; Cierpich, A.; Dudek, Ł.; Eggenweiler, H.-M.; Fogt, J.;

Gaik, M.; Gondela, A.; Jakubiec, K.; Jurzak, M.; Kitlińska, A.; Kowalczyk, P.; Kujawa, M.; Kwiecińska, K.; Leś, M.; Lindemann, R.; Maciuszek, M.; Mikulski, M.; Niedziejko, P.; Obara, A.; Pawlik, H.; Rzymski, T.; Sieprawska-Lupa, M.; Sowińska, M.; Szeremeta-Spisak, J.; Stachowicz, A.; Tomczyk, M. M.; Wiklik, K.; Włoszczak, Ł.; Ziemiańska, S.; Zarębski, A.; Krzysztof, B.; Nowak, M.; Fabritius, C.-H. Discovery and Structure–Activity Relationships of N-Aryl 6-Aminoquinoxalines as Potent PFKFB3 Kinase Inhibitors. *Chem. Med. Chem* **2019**, *14*, 169–181.

(71) Structure-activity Relationships of 2-Arylquinazolin-4-ones as Highly Selective and Potent Inhibitors of the Tankyrases. *Eur. J. Med. Chem.* **2016**, *118*, 316–327.

(72) Mcpherson, A. A Comparison of Salts for the Crystallization of Macromolecules. *Protein Sci.* **2001**, *10*, 418–422.

(73) Dessau, M. A.; Modis, Y. Protein Crystallization for X-ray Crystallography. *J. Vis. Exp.* **2011**, *47*, 2285.

(74) Buchstaller, H.-P.; Anlauf, U.; Dorsch, D.; Kuhn, D.; Lehmann, M.; Leuthner, B.; Musil, D.; Radtki, D.; Ritzert, C.; Rohdich, F.; Schneider, R.; Esdar, C. Discovery and Optimization of 2-Arylquinazolin-4-ones into a Potent and Selective Tankyrase Inhibitor Modulating Wnt Pathway Activity. *J. Med. Chem.* **2019**, *62*, 7897–7909.

(75) Hahn, D. F.; Gapsys, V.; de Groot, B. L.; Mobley, D. L.; Tresadern, G. J. Current State of Open Source Force Fields in Protein-ligand Binding Affinity Predictions. 2023, ChemRxiv; `https://chemrxiv.org/engage/chemrxiv/article-details/64e86ea479853bbd7862b98a`, (accessed October 22, 2023).

(76) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.

(77) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most

Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

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

(79) Horton, J. T.; Allen, A. E. A.; Dodda, L. S.; Cole, D. J. QUBEKit: Automating the Derivation of Force Field Parameters from Quantum Mechanics. *J. Chem. Inf. Model.* **2019**, *59*, 1366–1381.

(80) Galvelis, R.; Doerr, S.; Damas, J. a. M.; Harvey, M. J.; De Fabritiis, G. A Scalable Molecular Force Field Parameterization Method Based on Density Functional Theory and Quantum-Level Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 3485–3493.

(81) MolCharge Theory. `https://docs.eyesopen.com/applications/quacpac/theory/molcharge_theory.html`, (accessed April 28, 2023).

(82) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding: an Electrostatically-driven Highly Directional Noncovalent Interaction. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7748–7757.

(83) Jorgensen, W. L.; Schyman, P. Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents. *J. Chem. Theory Comput.* **2012**, *8*, 3895–3901.

(84) Parametrization of Halogen Bonds in the CHARMM General Force Field: Improved Treatment of Ligand–protein Interactions. *Bioorg. Med. Chem.* **2016**, *24*, 4812–4825.

(85) Nunes, R.; Vila-Viçosa, D.; Machuqueiro, M.; Costa, P. J. Biomolecular Simulations of Halogen Bonds with a GROMOS Force Field. *J. Chem. Theory Comput.* **2018**, *14*, 5383–5392.

(86) Yin, D.; MacKerell Jr, A. D. Combined ab initio/empirical Approach for Optimization of Lennard–Jones Parameters. *J. Comput. Chem.* **1998**, *19*, 334–348.

(87) Boothroyd, S.; Madin, O. C.; Mobley, D. L.; Wang, L.-P.; Chodera, J. D.; Shirts, M. R. Improving Force Field Accuracy by Training against Condensed-Phase Mixture Properties. *J. Chem. Theory Comput.* **2022**, *18*, 3577–3592.

(88) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312–2323.

(89) Boulanger, E.; Huang, L.; Rupakheti, C.; MacKerell Jr, A. D.; Roux, B. Optimized Lennard-Jones Parameters for Drug-like Small Molecules. *J. Chem. Theory Comput.* **2018**, *14*, 3121–3131.

(90) Kantonen, S. M.; Muddana, H. S.; Schauperl, M.; Henriksen, N. M.; Wang, L.-P.; Gilson, M. K. Data-Driven Mapping of Gas-Phase Quantum Calculations to General Force Field Lennard-Jones Parameters. *J. Chem. Theory Comput.* **2020**, *16*, 1115–1127.

TOC Graphic