

Implementation of the QUBE force field for high-throughput alchemical free energy calculations

Lauren Nelson,^{a†} Sofia Bariami,[‡] Chris Ringrose,[†] Joshua T. Horton,[†] Vadiraj
Kurdekar,[†] Antonia S. J. S. Mey,[‡] Julien Michel,[‡] and Daniel J. Cole^{*,†}

[†]*School of Natural and Environmental Sciences, Newcastle University, Newcastle upon
Tyne NE1 7RU, United Kingdom*

[‡]*EaStCHEM School of Chemistry, University of Edinburgh, David Brewster Road,
Edinburgh EH9 3FJ, United Kingdom*

E-mail: daniel.cole@ncl.ac.uk

^aL.N. and S.B. contributed equally to this work.

Abstract

The quantum mechanical bespoke (QUBE) force field approach has been developed to facilitate the automated derivation of potential energy function parameters for modelling protein-ligand binding. To date the approach has been validated in the context of Monte Carlo simulations of protein-ligand complexes. We describe here the implementation of the QUBE force field in the alchemical free energy calculation molecular dynamics simulation package SOMD. The implementation is validated by computing relative binding free energies for two congeneric series of non-nucleoside inhibitors of HIV-1 reverse transcriptase using QUBE and AMBER/GAFF force fields. The availability of QUBE in a modern simulation package that makes efficient use of GPU acceleration will greatly facilitate future high-throughput alchemical free energy calculation studies.

Introduction

The ability to accurately predict protein-ligand binding affinity is invaluable in the early stages of drug discovery. High-throughput alchemical free energy calculations are an attractive tool for this task, enabling rigorous calculation of binding free energies. However, accuracy remains limited by the description of interatomic interactions and the sampling of conformational space.¹⁻⁴ The potential energy surfaces of protein-ligand complexes are almost always described by molecular mechanics (MM) force fields, of which AMBER,⁵ OPLS,⁶ CHARMM⁷ and GROMOS⁸ are popular examples. These transferable biological force fields all employ similar functional forms and their parameters are typically fit to reproduce the quantum mechanical (QM) and/or experimental properties of small organic molecules.

Recently, an alternative approach to biomolecular force field design has been proposed, named the QUantum mechanical BEspoke (QUBE) force field, in which virtually all potential parameters are derived specifically for the molecule under study directly from a small number of QM calculations. QUBE shares its functional form with the OPLS force field, and so is rapid to evaluate in the context of alchemical free energy calculations. Full details may be found elsewhere.^{9,10} In brief, non-bonded (charge and Lennard-Jones) parameters of the QUBE force field are derived from atoms-in-molecule partitioning of the ground state QM electron density,^{11,12} in particular, employing the Tkatchenko-Scheffler relations for van der Waals interactions.¹³ QUBE bond and angle parameters are derived from the QM Hessian matrix, using the modified Seminario method,^{14,15} while anharmonic dihedral parameters are fit to relaxed QM torsion scans.⁹ Small molecule QUBE force fields may be derived using the QUBEKit python package, and they have been extensively validated against experimental liquid properties.⁹ Atoms-in-molecule protocols are available as part of the ONETEP linear-scaling density functional theory software,¹⁶ and hence QUBE non-bonded parameters may be readily derived for entire proteins comprising thousands of atoms. These parameters have been supplemented by compatible libraries of bonded parameters, and simulations of protein dynamics using the resulting QUBE force fields have been validated against experimental

NMR observables.¹⁰

To date, the QUBE force field has been used to compute absolute binding free energies of a series of benzene derivatives to the L99A mutant of T4 lysozyme¹⁷ and relative binding free energies of several flexible inhibitors of p38 α MAP kinase.¹⁸ In both cases, mean unsigned errors (MUEs) in protein-ligand binding free energies of under 1 kcal/mol were reported. Such accuracy was shown to be broadly similar to that of the widely-used OPLS force field in these cases. In the initial development of QUBE, the OPLS functional form has been retained for compatibility with existing MM software (for example, these studies used the MCPRO software¹⁹). However, now that a baseline accuracy has been established, future development strategies will target rapid and systematic evolution of the force field functional form. Such a strategy involves identifying key mappings between QM observables (such as the electron density) and force field parameters. Examples include the automated addition of off-center charges to account for anisotropic electron density,⁹ which has not yet been thoroughly explored in protein-ligand binding studies,¹¹ and higher-order dispersion terms to move beyond the dipole-dipole r^{-6} interaction.^{20,21}

For rapid testing of these new force fields, it is desirable to interface with free energy software that can be readily adapted to new force field functional forms whilst achieving efficient performance on modern computing hardware. Options for high-throughput alchemical free energy simulations include Schrödinger’s commercial FEP+ package,²² as well as AMBER⁵ and GROMACS.⁸ However, packages such as these tend to be developed specifically for a given force field and file format, reducing interoperability and ease of comparison. Indeed, a recent study of relative hydration free energies using different simulation packages, highlighted the careful steps required to reproduce this quantity across simulation packages.²³ For this reason, here we implement QUBE in the Sire molecular simulation framework²⁴ which includes the SOMD molecular dynamics engine for free energy calculations. SOMD interfaces with the OpenMM toolkit for GPU acceleration.²⁵ We also make use of the BioSimSpace library which facilitates system setup and interoperability between a range of biomolecular

simulation packages, such as AMBER, GROMACS, and CHARMM, to facilitate preparation of QUBE inputs.²⁶ SOMD has been used within the Sire molecular simulation framework and successfully applied to alchemical free energy studies on a range of drug-like fragments, carbohydrates and host-guest systems.²⁷⁻³⁷

The automated setup and processing of alchemical free energy calculations using Sire, BioSimSpace, SOMD and OpenMM has recently been implemented in Cresset’s Flare package,³⁸ and benchmarked against 220 ligands bound to 14 protein targets, with accuracy comparable to previous reports.^{3,8,22} Thus, the SOMD framework provides a promising basis for implementation and future benchmarking of alchemical transformations using the QUBE force field. In this paper, we develop and distribute the file parsers that allow users to run QUBE simulations of protein-ligand complexes in the Sire and OpenMM molecular simulation frameworks. We further demonstrate the use of the QUBE force field in alchemical free energy simulations with focus on datasets created from extensive work on the HIV-1 reverse transcriptase (HIV-1 RT) protein target by the Jorgensen group.³⁹⁻⁴³

HIV-1 is one of two strains of HIV which cause AIDS; the most advanced stage of the HIV infection. The rapid replication of HIV-1 can lead to errors in viral replication and virus evolution, causing difficulties for effective vaccine development.⁴⁴ There are three viral enzymes that are targeted for drug development, one of which is HIV-1 reverse transcriptase (RT). HIV-1 RT is an asymmetric heterodimer with two subunits: p66, the larger of the two subunits, which contains the active sites of the two enzymatic functions of RT, and p51, the smaller subunit which has a structural role.⁴⁴ HIV-1 relies on RT to copy its single-stranded RNA genome into a double stranded DNA copy before it can be integrated into the host cell genome; preventing this reduces viral replication. RT first binds to DNA or RNA, followed by the binding of deoxynucleoside triphosphate (dNTP) in a two-step process, to form a tertiary complex. This induces a conformational change, enabling the 3’-hydroxyl of the elongating strand to attack the α -phosphate of dNTP, before the elongated DNA is released from RT.⁴⁵

The molecules investigated here are potential non-nucleoside inhibitors of HIV-1 RT (NNRTIs), which block polymerase activity of RT.^{39–42} NNRTIs are non-competitive inhibitors, binding to a hydrophobic allosteric pocket which is adjacent to the polymerase active site, causing a conformation change in the enzyme, thus preventing the substrate from binding.⁴⁴ The allosteric site consists of 15 residues in the p66 subunit (L100, K101, K103, V106, T107, V108, V179, Y181, Y188, V189, G190, F227, W229, L234, and Y318) and one residue in the p51 subunit (E138). As shown in Figure 1(A), the binding pocket accommodates well-packed aryl-aryl interactions with Tyr181, Tyr188, and Trp229.⁴⁰ There is room within the hydrophobic allosteric site for potential NNRTIs to venture toward Pro95, located above Tyr181 and below Trp229.

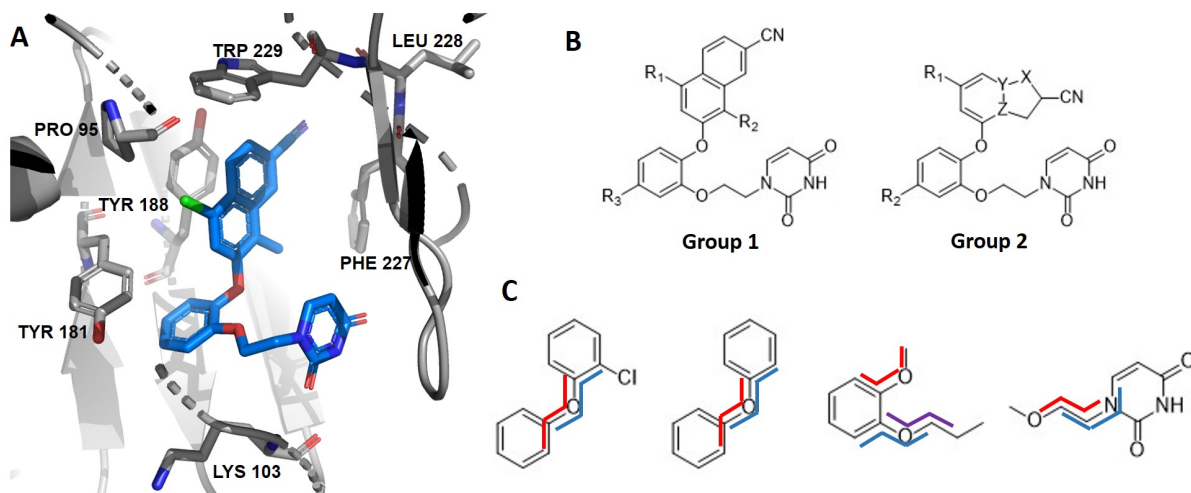


Figure 1: (A) Example X-ray crystal structure of the HIV-1 RT NNRTI binding site (PDB: 5ter) with a bound catechol diether inhibitor, and key amino acids highlighted.⁴⁰ (B) Markush structures of group 1 (2-naphthyl analogs) and group 2 (indole, indolizine and benzofuran analogs) compounds.^{39,40} (C) The fragments used to generate torsional parameters, with the dihedral angles that were fit for each fragment highlighted.

Jorgensen and co-workers have worked on the HIV-1 complex for over 20 years, developing bicyclic NNRTIs targeting HIV-1 RT; initially pursuing diarylamines and pyrimidinyl- and triazinyl-amines.^{40,41,46,47} In an effort to replace reactive cyanovinyl groups present in early designs, the group used free energy perturbation (FEP) calculations to scan viable scaffolds, and found indoles, indolizines and benzofurans to be the most viable options (Figure 1(B),

group 2).³⁹ Synthetic studies and assays were conducted, and structural characterisation with X-ray crystallography confirmed their findings, showing good positioning in the allosteric binding site. This work uncovered novel, potent NNRTI agents that displayed low cytotoxicity and good solubility.

Following their research with indolizine variants,³⁹ Jorgensen and co-workers investigated bicyclic replacements to overcome lower potency towards the Y181C viral mutation, subsequently focusing on naphthyl analogues.⁴⁰ Using X-ray crystal structures for various catechol diethers bound to wild type (WT) and other HIV-1 RT variants, the group used *de novo* design and free energy calculations⁴⁰ to build alternative 2-naphthyl compounds (Figure 1(B), group 1). Despite their larger size, these are accommodated by the NNRTI binding site and are viable if the cyano group is in the 7-position. This work subsequently generated several analogs for synthesis, with good potency against both the WT virus, and mutant strains.

Here, we chose 10 compounds from the 2-naphthyl dataset⁴⁰ (group 1, Table 1) and 12 compounds comprising indoles, indolizines, and benzofurans from earlier work³⁹ (group 2, Table 2). We focused on these two datasets as group 1 exemplifies small side chain transformations, whilst group 2 covers a much wider potency range, including heterocyclic substitutions that are expected to provide a challenge for force field design. The general structure of each group is displayed in Figure 1(B). For each dataset we can compare computational findings with experimental inhibitory activity measured in MT-2 cell assays, which are presented in Tables 1 and 2.^{39,40}

In what follows, we report the derivation of force field parameters for the protein and small molecules with the QUBEKit software, and describe our new interface with the SOMD package. We compute relative binding free energies for each of the inhibitors to the HIV-1 RT target, described above, and compare our data with experiment and with the widely used GAFF2 and AMBER force fields.

Table 1: Group 1 experimental EC₅₀ inhibitory activity in WT HIV-1 RT assays.⁴⁰

Compound	R_1	R_2	R_3	EC ₅₀ (nM)
1a	F	Cl	H	5.0
1b	F	Me	H	7.8
1c	Cl	Me	H	6.2
1d	Me	Cl	H	5.0
1e	Me	Me	H	3.5
1f	Et	Me	H	6.0
1g	Pr	Me	H	21.0
1h	iPr	Me	H	16.0
1i	Me	F	F	58.0
1j	Me	Me	F	1.9

Table 2: Group 2 experimental EC₅₀ inhibitory activity in WT HIV-1 RT assays.³⁹

Compound	R_1	R_2	X	Y	Z	EC ₅₀ (nM)
2a	H	H	-	N	-	0.38
2b	Me	H	-	N	-	0.9
2c	Me	F	-	N	-	2.0
2d	F	F	-	N	-	0.4
2e	H	F	-	N	-	2.7
2f	H	Cl	-	N	-	5.1
2g	H	F	-	-	N	17
2h	H	F	O	-	-	40
2i	Me	F	O	-	-	260
2j	H	H	NH	-	-	56
2k	Me	H	NH	-	-	10
2m	Cl	H	NH	-	-	340

Computational Methods

Protein and Ligand Setup

Simulations for groups 1 and 2 were based on the initial crystal structures with PDB codes 5ter and 4mfb, respectively. The protein was pre-processed using the visualization software Maestro.⁴⁸ All residues further than 15 Å from the ligand binding site were removed, and dangling bonds were capped using NME and ACE groups. The final system sizes were 376 and 384 residues for groups 1 and 2, respectively. The protein underwent a short minimization and equilibration with backbone atoms restrained. Starting from the ligands in the two crystal structures (that is, compound **1c** for group 1, and compound **2a** for group 2), initial structures for all 22 ligands were built, and hydrogens added using Maestro.

For direct comparison with QUBE on this data set, we employ the widely-used Amberff14sb force field for the protein and GAFF2 for the ligands (hereafter referred to as AMBER). BioSimSpace was used to parameterize the proteins and ligands, generating AMBER/GAFF2 parameters and their charges whilst utilizing the AM1-BCC charge model for the ligands. Each ligand was combined with the corresponding protein, and both unbound ligands and the complexes were solvated using the TIP3P water model with cubic box sizes of 26 Å and 88 Å respectively. Short equilibration simulations were run using BioSimSpace for every solvated system to produce files ready for use in free energy calculations.

Ligand Preparation with QUBEKit

QUBEKit⁹ interfaces with the Gaussian09⁴⁹ and ONETEP¹⁶ QM software packages to perform bond, angle, torsion, charge and Lennard-Jones parameter derivation. BOSS/MCPRO style z-matrices, and the corresponding PDB files, of the ligands were generated using the LigParGen⁵⁰ web server. Gaussian09 input files were prepared using QUBEKit. Structural optimizations and Hessian matrix calculations were performed with the ω B97X-D⁵¹ functional and a 6-311++G(d,p) basis set. Harmonic bond stretching and angle bending pa-

rameters were derived using the modified Seminario method,^{14,15} with a vibrational scaling factor of 0.957.¹⁵

Non-bonded parameter derivation was performed with the linear-scaling DFT code, ONETEP,¹⁶ using previously reported protocols.⁹ The ground state electron density was computed using an implicit solvent model with a dielectric of 4 to model induction effects in an effective manner.¹¹ The DDEC module implemented in ONETEP was used to partition the electron density and assign atom-centered point charges and atomic volumes.^{9,12,52,53} Partitioning was performed with an IH to ISA ratio of 0.02. Lennard-Jones parameters were derived by QUBEKit from the atomic volumes using the Tkatchenko-Scheffler method.^{11,13} No off-site charges were used in this work.⁹

Torsion parameter fitting follows the general methods used in previous studies,^{9,18} though here we employ an interface between QUBEKit and the TorsionDrive package,⁵⁴ which improves the quality of both QM and MM torsion scans through its recursive wavefront propagation algorithm. QM torsion scans use the same functional basis set as used to derive the bond and angle parameters above. To aid QM optimization and save computational expense, torsional scans were conducted on core fragments of the ligand sets (Figure 1(C)). The iterative fitting algorithm is described in the **Supporting Methods**. Non-flexible torsional parameters (*e.g.* for ring systems) were taken from the OPLS force field as provided by the LigParGen⁵⁰ web server. For group 2, torsions involving the cyano group (N-C-C-C) were assigned erroneous atom types by LigParGen, and so these parameters were manually set to zero. Final force fields were output in xml format. A full list of commands for QUBEKit can be found on our Github page (<https://github.com/qubekit/QUBEKit>).

Protein Preparation with QUBEKit and Sire

The truncated protein target underwent parameterization using the QUBE force field. ONETEP calculations were performed to obtain the ground state electron density, atomic charges and atomic volumes, in the same way as the ligands, as previously described.¹⁸

To facilitate set-up of QUBE simulations for proteins, the software QUBEKit-pro has been developed and is used here to generate OpenMM xml files from pdb and ONETEP output files. Using QUBEKit as a base has the advantage that most features can be easily applied to the protein, allowing for charge checking and symmetrization, for example. QUBEKit-pro builds the xml by first reading the PDB file to obtain the full topology of the protein or fragment. This provides atomic positions and connectivity, as well as some other information, such as the amino acid sequence and whether the protein is split into subunits. At this stage, certain groups of atoms are picked out for later symmetrization such as hydrogen atoms on the same methyl or amine groups.

With the structure read in, a parametrization step is performed. A stored, general protein xml file contains bond, angle and torsion parameters, which have been generated specifically for compatibility with the QUBE force field¹⁰ and are available for all atoms in standard residues, including NME/ACE caps. This general xml is used to map parameters to the protein in question, using the now stored structure. As described above for the ligands, atomic charges and volumes are extracted from the ONETEP output file, symmetrized (if required), and used to calculate the Lennard-Jones parameters.¹¹

QUBEKit is then used to write final pdb and xml files for use with OpenMM. If the protein consists of multiple chains, each molecule is returned as a separate file. Since each atom in the protein is in a unique environment, and therefore has unique charge and Lennard-Jones parameters, each atom in QUBEKit-pro is assigned a unique type.

Molecular systems in Sire can be created using AMBER style input files. The molecule is described primarily with two files: the prmtop (or prm7) file that holds all the parameters and the inpcrd (or rst7) that contains the coordinates of the atoms. These files are generated during the parameterization of the molecule of interest with the AMBER force field. Given that all the molecules are now parameterized with the QUBE force field, functionality in Sire was extended to include new features that support parsing of the new pdb and xml input files and support the OPLS-type potential energy functional form. For the former,

an algorithm was implemented that reads the parameters and the coordinates from the xml and pdb input files respectively, and returns formatted AMBER files (prm7/rst7) that contain exactly the same information and can be parsed by SOMD. Geometric combination rules have also been implemented in SOMD to support OPLS-style force fields. Further information and benchmarking of single point energies in OpenMM and SOMD are provided in the **Supporting Methods**.

Free Energy Simulations

BioSimSpace was used to generate files for free energy calculations of the ligands in bound and unbound states. Free energy simulations were run with SOMD using both AMBER and QUBE force field parameterization methods for comparison. Perturbation maps for each HIV-1 RT data set were constructed manually (**Figure S3**). For the 10 ligands used in the group 1 data set the free energy map had 32 transitions, whilst the 12 group 2 ligands comprised 42 transitions. In both cases, multiple cycle closures allow assessment of the convergence of the computed free energies.

Each perturbation was carried out in two independent simulations, one in each direction. Though this increased the overall computational cost, it provided an opportunity to assess the precision of the calculations and check for hysteresis. Each bound and unbound simulation was divided into 11 regularly spaced λ windows. The energy of the system was minimized for 1000 cycles by using the steepest descent method and each λ window was run for a total of 4 ns, using a time step of 2 fs. The first 5% of each simulation trajectory was discarded as equilibration. **Table S2** shows that the transformation of **2a** to **2c** (and the reverse) in group 2 varies by less than 0.3 kcal/mol with respect to changes in simulation time (in the range 2–4 ns).

Bonds involving hydrogen, that were not involved in an alchemical transformation, were constrained to their equilibrium distances. Periodic boundary conditions and a 10 Å cut-off distance for non-bonded interactions were applied. In the solvent simulations, an Andersen

thermostat with collision frequency 10.0 ps^{-1} and a Monte Carlo barostat with a frequency of 25 time steps were used to keep the temperature and the pressure constant. Finally, the electrostatic interactions were calculated using reaction field with the water dielectric constant (78.3).

Free energy changes were calculated from the output using MBAR and thermodynamic integration (TI) methods, as implemented in pymbar which is integrated into the Sire application `analyse_freenrg`. Only MBAR results are presented here for simplicity, though TI data were checked to be similar (see **Table S2**, for example). The collection of MBAR relative binding free energies was then processed by the software FreeEnergyWorkflows²⁷ to produce the free energy estimates, and associated errors, reported in this manuscript.

All parameterization and free energy protocols, example input files and output data may be found at <insert github link>.

Results

We begin analysis with group 1, the catechol diether that incorporates a 7-cyano-2-naphthyl substituent. Figure 2(A) shows an overlay of compound **1c** from MD simulations using the QUBE force field, with the corresponding crystal structure. As expected, the cyano group projects out below Trp229 into a solvent-exposed channel. The 2-naphthyl group maintains close aryl-aryl contacts with Tyr188 and Trp229.⁴⁰ Tyr181 has switched orientation from a face-to-face interaction with the catechol diether ring of the ligand, to form closer contact with the 2-naphthyl group. However, **Figure S4** reveals that Tyr181 is very flexible, forming both edge-to-face and face-to-face interactions throughout the simulation that more closely resemble the crystal structure. The observed flexibility of Tyr181 is consistent with previous simulations⁵⁵ and crystallography,³⁹ which suggest that Tyr181 orientation is sensitive to the identity of the ligand. One of the carbonyl oxygens of the uracil group maintains hydrogen bonding interactions with the backbone of Lys103 throughout the MD simulations, while the

nitrogen and second carbonyl remain solvent-exposed. The orientation of the ethoxy linker connecting the uracil and catechol diether rings is different to that observed in the crystal structure. However, again, MD trajectories sample multiple conformations over the time scale of these simulations (**Figure S4**), and crystal structures of similar catechol diethers support the feasibility of the alternate conformation.⁵⁶

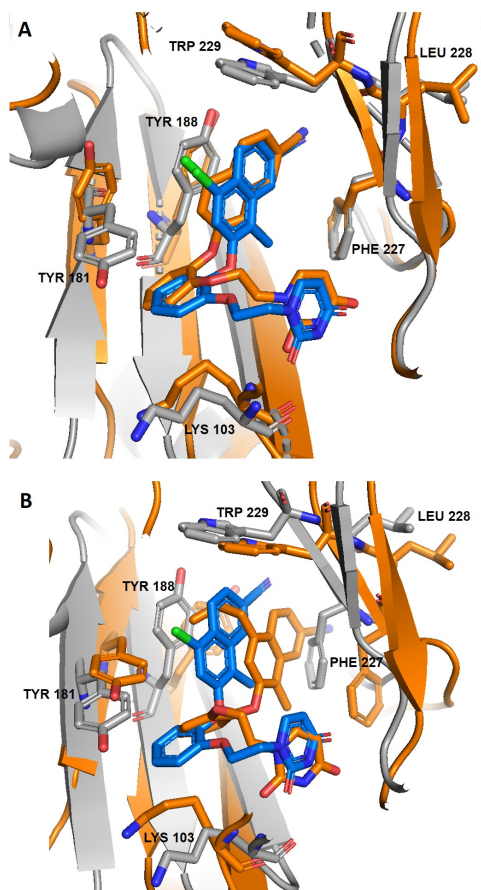


Figure 2: Overlay of (A) compound **1c** and (B) compound **1g** (orange) from group 1 with the crystal structure of **1c** (PDB: 5ter, grey/blue).

Figure 3 displays the binding affinities of the 2-naphthyl ethers (relative to compound **1e**), computed using the QUBE force field and compared with experimental assays, using human T-cells infected by wild-type HIV-1 (MT-2 cell assays),⁴⁰ plotted as differences in pEC_{50} . For direct comparison, we also report relative binding free energies computed with the AMBER force field, using the same computational set-up. Compounds **1a–1e** differ from each other through a series of small methyl and halogen transformations at the R_1 and R_2 positions

(Table 1). With $R_2 = \text{Me}$, the Me and Cl substitutions at R_1 are favored over F (**1e** < **1c** < **1b**), an ordering which is recapitulated by QUBE while AMBER strongly favors compound **1c**. Experimentally, the two compounds (**1a** and **1d**) with $R_2 = \text{Cl}$ are equipotent and less strongly bound than **1e**. In this case, QUBE obtains the correct sign of the transformation for **1a**, but slightly favors compound **1d**.

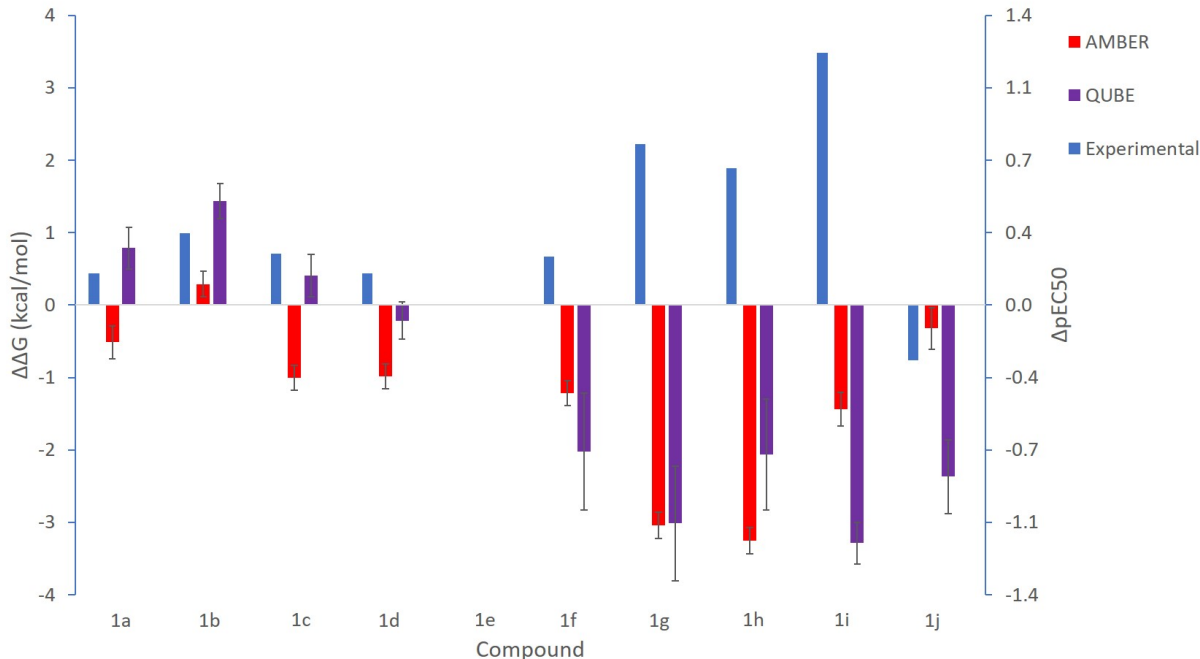


Figure 3: Binding free energies (relative to compound **1e**) computed using the QUBE and AMBER force fields (left axis), and compared with differences in experimental pEC_{50} assay results (right axis).⁴⁰

In contrast to the promising results above, the two compounds (**1i** and **1j**) with $R_3 = \text{F}$ are strongly over-bound with the QUBE force field. The small potency gain expected from the $R_3 = \text{H}$ to F transformation (**1e** to **1j**) is modelled well with AMBER, but over-estimated by QUBE by around 2 kcal/mol. **Figure S5** shows significant movement of the catechol diether group of compound **1j** towards the space between Tyr181 and Lys103, which represents a larger change in conformation than was observed for the other inhibitors. Compound **1i** is expected to be the most weakly bound of the 2-naphthyl ethers investigated here, but is strongly over-bound by both AMBER and QUBE. Interestingly, we did not

include a $R_2 = F$ substituent in our fragment torsion scans (Figure 1(C)), and it may be that the electronegative F atom over-stabilizes the bound conformation in combination with both the QUBE force field parameters and the transferable GAFF library. It is also worth noting that the 2-naphthyl ether, with $R_1 = R_2 = H$, has been shown to adopt an alternative binding mode to the one shown in Figure 2.⁴⁰ Compound **1i** could be thus expected to pose a challenge if the small $R_2 = F$ substituent causes a large conformational change in the ligand. **Table S3** reports a series of repeat runs for several transformations involving compounds **1i** and **1j**. We do not observe very large differences between repeated runs, but do see large hysteresis between forward/reverse transformations (close to 2 kcal/mol in some cases), which is again indicative of sampling inadequacies.

It has been hypothesized that bulkier groups at the R_1 position in the 2-naphthyl ethers might be accommodated, and confer extra benefit in the common Y181C mutant viral strain of HIV-1 RT.⁴⁰ Hence, compounds **1f–1h** were added to our benchmark data set. Both AMBER and QUBE force fields strongly over-estimate binding in these cases. There is evidence of a change in binding mode of these bulkier ligands (Figure 2(B)), particularly for **1g** and **1h**. Alongside the large error bars on these computed data and the insensitivity to the choice of force field, this indicates that insufficient sampling may be an issue here. Although we have investigated typical convergence of the free energies over simulation time scales of 2–4 ns (**Table S2**), it may be that significantly longer simulation times and/or enhanced sampling are required here.⁵⁷

We turn now to group 2, where the 2-naphthyl group is replaced by a series of indolizines, indoles, and furans, which exemplify the challenge of optimizing heterocycles in early stage drug discovery. A crystal structure is available for compound **2a**,³⁹ and this is shown in Figure 4, overlaid with the final frame from our MD simulations with the QUBE force field. The main interactions formed are similar to those discussed above. Tyr188 and Thr229 form aryl-aryl interactions with the indolizine, and the single uracil hydrogen bond with Lys103 is retained. In this case, Tyr181 adopts a T-shaped stacking interaction with the central

catechol ring. In MD simulations, both the orientation of Tyr181 and the conformation of the ethoxy linker are stable and in agreement with the X-ray crystal structure. Some rotation of the indolizine group is visible, but again this is flexible throughout the MD simulation (Figure S6).

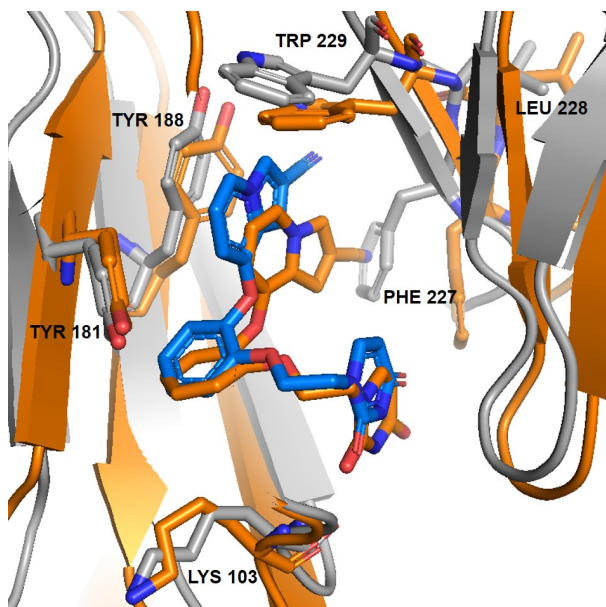


Figure 4: Overlay between compound **2a** of group 2 (orange) and the crystal structure (PDB: 4mfb, grey/blue).

As before, we have computed the relative binding affinities of each of the group 2 compounds in Table 2 with both the AMBER and QUBE force fields, and compared with available assays.³⁹ Figure 5 reveals very good agreement with experimental trends across all compounds for both force fields. Compounds **2a** to **2f** represent a series of indolizines, with methyl and halogen substitutions in the R₁ and R₂ positions. Compounds **2a**, **2b** and **2d** are all sub-nanomolar inhibitors, and are predicted by our simulations using the QUBE force field to be essentially equipotent. Although AMBER predicts the correct sign for the transformation of **2c** to **2d**, it does underestimate the potency of **2a** and **2b** (to the extent that they may not have been synthesized in a prospective study). Both force fields correctly describe the drop in affinity of compounds **2e** and **2f**. It is interesting that **2c–2f**, with halogen substitutions at the R₂ position, are modelled well with QUBE which contrasts with the

issues observed for halogen substitutions at R₃ in group 1. This may point to a lack of space in which to accommodate the F substitution in the latter, and possible associated sampling issues. The isomeric indolizine (**2g**) was included in our test set to analyze the effects of more subtle electrostatic aryl-aryl interactions on binding. Both QUBE and AMBER recapitulate the expected drop in potency, though QUBE (in agreement with Monte Carlo simulations performed with the OPLS force field³⁹) likely over-estimates the magnitude of this change.

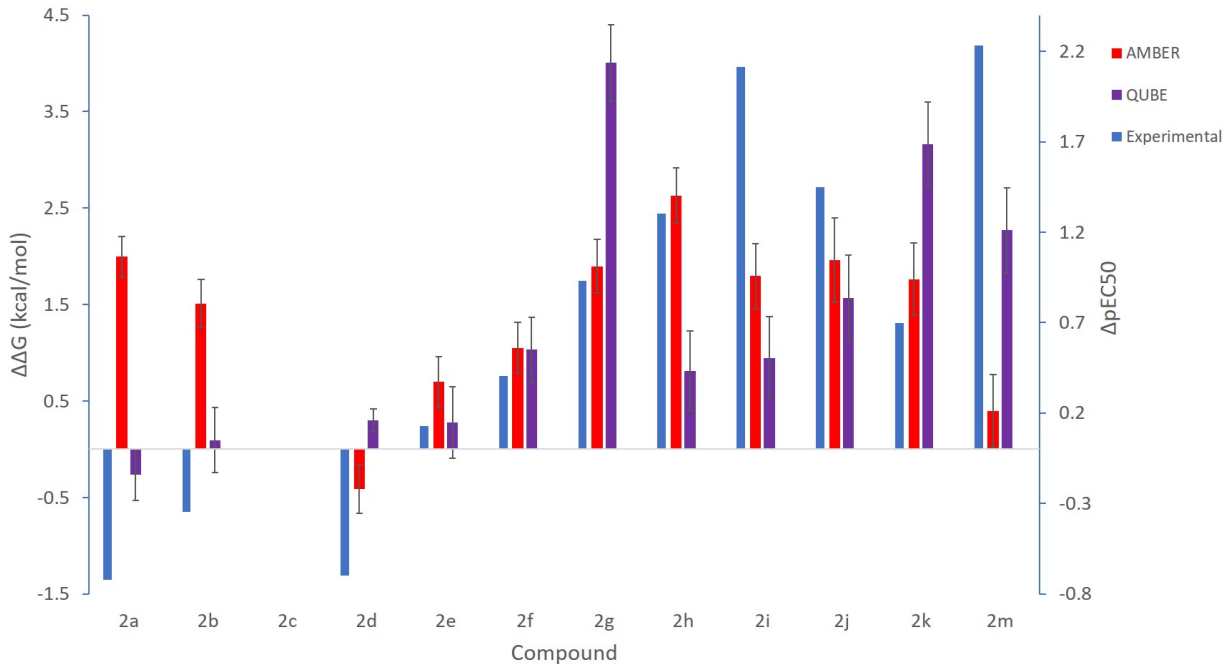


Figure 5: Binding free energies (relative to compound **2c**) computed using the QUBE and AMBER force fields (left axis), and compared with differences in experimental pEC₅₀ assay results (right axis).³⁹

Finally, series of benzofurans (**2h** and **2i**) and indoles (**2j–2m**) were investigated. Both force fields successfully recover the drop in potency (relative to the indolizines), which is encouraging for future prospective heterocycle scans. QUBE slightly over-estimates the affinity of the two furans, and under-estimates that of the indole with a methyl substitution at R₁ (**2k**), relative to the other indoles, but is in otherwise very good agreement with experimental trends.

Discussion and Conclusions

In this paper, we have described an interface between the QUBEKit force field engine, and the Sire molecular simulation framework, for the calculation of alchemical relative binding free energies. As an example application, we have retrospectively analyzed the binding of 22 small molecule allosteric inhibitors of HIV-1 reverse transcriptase. As discussed earlier, both AMBER and QUBE perform well for small substitutions on the 2-naphthyl of group 1, but uniformly over-bind compounds with bulkier hydrocarbon chains and (to some extent) those with halogens in the R_3 position. With the significant caveat that EC_{50} is not a direct measurement of binding, we can estimate experimental binding free energies from the available data. The mean unsigned error in QUBE predictions, relative to experiment, is then 1.85 kcal/mol, which is slightly higher than AMBER (1.66 kcal/mol). It has been shown previously, in the context of Monte Carlo simulations, that enhanced sampling is required to mitigate the effects of quasi-ergodic sampling for some bulky inhibitors in this particular binding site.⁵⁸ We propose that the present group 1 set may represent an interesting test case for such sampling protocols.

In contrast, the group 2 set seems to represent a ‘well-behaved’ test case for free energy methods. Overall, compared with group 1, the conformational dynamics of group 2 compounds retain much stronger agreement with the available crystal structure, which may help to explain the relatively better performance of both QUBE and AMBER on this set. Relative to experiment, the mean unsigned errors are 1.05 and 1.08 kcal/mol for the QUBE and AMBER force fields, respectively. In fact, we may also compare these results with the CHARMM36/CGenFF force field, which was employed as part of a λ -dynamics protocol by Vilseck et al.⁵⁵ For similar indole and indolizine test sets, they report mean unsigned errors of 1.13 and 0.70 kcal/mol, respectively, though note that they did not directly transform between these scaffolds as we have done here. Thus, for both transferable (AMBER and CHARMM) and bespoke (QUBE) force field parameter sets, the errors reported on this set are in agreement with the current consensus on the accuracy of free energy methods in

computer-aided drug design.^{8,22,38}

Despite the relatively similar overall performance of AMBER and QUBE on these datasets, **Figure S7** shows that there is little similarity between predictions of individual relative binding free energies using the two force fields. This is perhaps to be expected given the fundamentally very different approaches to parameter derivation. While, at first sight, there is no particular reason to use QUBE over, for example, the AMBER or CHARMM force fields at this time, our bespoke parameter derivation methodologies do offer the potential for substantial improvements in accuracy down the line. By deriving as many force fields parameters as practical directly from QM, rather than fitting to experiment, the process of force field design becomes an exercise in accurately mapping QM data onto physically-motivated parameters and functional forms.^{9,59,60} We emphasize that in developing the QUBE non-bonded parameters for protein-ligand complexes such as these, only seven parameters have been directly fit to experiment (the van der Waals radii of seven elements), with the remainder being derived from QM. Thus, future parameter and functional form updates should be relatively straightforward to automate. In contrast, similar adjustments of the functional forms of the aforementioned transferable force fields would be a substantial undertaking, requiring a complete re-fit of all force field parameters to experimental data (admittedly a process that is gradually becoming more achievable through efforts such as the Open Force Field Initiative⁶¹). As an example, a polarizable force field, supplemented by higher order (C_8) dispersion coefficients derived from QM, has recently been shown to perform very well in liquid simulations of alkanes.²¹ Indeed, it has been shown that such parameters, or force field precursors, may be derived from the same atoms-in-molecule technique used by QUBEKit to parameterize force fields.^{20,62} Overall, this first generation interface between QUBE and SOMD offers a robust, adaptable platform, with access to GPU-accelerated dynamics, that will substantially improve our ability to validate and apply bespoke QM-derived force fields in computer-aided drug design.

Acknowledgement

This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (<http://www.hpc.cam.ac.uk>) funded by EPSRC Tier-2 capital grant EP/P020259/1, as well as the Rocket High Performance Computing service at Newcastle University. DJC and VK acknowledge financial support from EPSRC grant EP/R010153/1.

Supporting Information Available

Supporting methods, free energy mappings, supporting MD snapshots, convergence data, and QUBE/AMBER comparison (PDF).

References

- (1) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531–558.
- (2) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.
- (3) Song, L. F.; Merz, K. M. Evolution of Alchemical Free Energy Methods in Drug Discovery. *J. Chem. Inf. Model.* **2020**,
- (4) Mey, A. S. J. S.; Allen, B.; Macdonald, H. E.; Chodera, J. D.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations. *arXiv:2008.03067* **2020**,

- (5) 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.
- (6) 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.
- (7) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell Jr., A. CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comp. Chem.* **2010**, *31*, 671–690.
- (8) Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; Van Vlijmen, H.; Tresadern, G.; De Groot, B. L. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **2020**, *11*, 1140–1152.
- (9) Horton, J. T.; Allen, A. E.; 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.
- (10) Allen, A. E.; Robertson, M. J.; Payne, M. C.; Cole, D. J. Development and Validation of the Quantum Mechanical Bespoke Protein Force Field. *ACS Omega* **2019**, *4*, 14537–14550.
- (11) 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.
- (12) Manz, T. A.; Sholl, D. S. Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials. *J. Chem. Theory Comput.* **2012**, *8*, 2844–2867.

- (13) Tkatchenko, A.; Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **2009**, *102*, 6–9.
- (14) Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **1996**, *60*, 1271–1277.
- (15) 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.
- (16) Prentice, J. C. et al. The ONETEP linear-scaling density functional theory program. *J. Chem. Phys.* **2020**, *152*, 174111.
- (17) Cole, D. J.; Cabeza De Vaca, I.; Jorgensen, W. L. Computation of protein-ligand binding free energies using quantum mechanical bespoke force fields. *MedChemComm* **2019**, *10*, 1116–1120.
- (18) Horton, J. T.; Allen, A. E.; Cole, D. J. Modelling flexible protein-ligand binding in p38 α MAP kinase using the QUBE force field. *ChemComm* **2020**, *56*, 932–935.
- (19) Jorgensen, W. L.; Tirado-Rives, J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comp. Chem.* **2005**, *26*, 1689–1700.
- (20) Manz, T.; Chen, T.; Cole, D. J.; Limas, N. G.; Fiszbein, B. New scaling relations to compute atom-in-material polarizabilities and dispersion coefficients: part 1. Theory and accuracy. *RSC Adv.* **2019**, *9*, 19297–19324.
- (21) Visscher, K. M.; Geerke, D. P. Deriving Force-Field Parameters from First Principles Using a Polarizable and Higher Order Dispersion Model. *J. Chem. Theory Comput.* **2019**, *15*, 1875–1883.

- (22) Wang, L. et al. 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.
- (23) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14*, 5567–5582.
- (24) Sire Molecular Simulation Framework. 2019; <http://siremol.org>.
- (25) 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*, 1–17.
- (26) Hedges, L.; Mey, A.; Laughton, C.; Gervasio, F.; Mulholland, A.; Woods, C.; Michel, J. BioSimSpace: An interoperable Python framework for biomolecular simulation. *J. Open Source Softw.* **2019**, *4*, 1831.
- (27) Mey, A. S.; Jiménez, J. J.; Michel, J. Impact of domain knowledge on blinded predictions of binding energies by alchemical free energy calculations. *J. Comput. Aided Mol. Des.* **2018**, *32*, 199–210.
- (28) Bosisio, S.; Mey, A. S.; Michel, J. Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *J. Comput. Aided Mol. Des.* **2017**, *31*, 61–70.
- (29) Granadino-Roldán, J. M.; Mey, A. S.; Pérez González, J. J.; Bosisio, S.; Rubio-Martinez, J.; Michel, J. Effect of set up protocols on the accuracy of alchemical free energy calculation over a set of ACK1 inhibitors. *PLoS ONE* **2019**, *14*, 1–20.

- (30) De Simone, A.; Georgiou, C.; Ioannidis, H.; Gupta, J., A. A. Juárez-Jiménez; Doughty-Shenton, D.; Blackburn, E. A.; Wear, M. A.; Richards, J. P.; Barlow, P. N.; Carragher, N.; Walkinshaw, M. D.; Hulme, A. N.; Michel, J. A computationally designed binding mode flip leads to a novel class of potent tri-vector cyclophilin inhibitors. *Chem. Sci.* **2019**, *10*, 542–547.
- (31) Papadourakis, M.; Bosisio, S.; Michel, J. Blinded predictions of standard binding free energies: lessons learned from the SAMPL6 challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1047–1058.
- (32) Rodil, A.; Bosisio, S.; Ayoup, M. S.; Quinn, L.; Cordes, D. B.; Slawin, A. M.; Murphy, C. D.; Michel, J.; O’Hagan, D. Metabolism and hydrophilicity of the polarised ‘Janus face’ all-*cis* tetrafluorocyclohexyl ring, a candidate motif for drug discovery. *Chem. Sci.* **2018**, *9*, 3023–3028.
- (33) Georgiou, C.; McNae, I.; Wear, M.; Ioannidis, H.; Michel, J.; Walkinshaw, M. Pushing the Limits of Detection of Weak Binding Using Fragment-Based Drug Discovery: Identification of New Cyclophilin Binders. *J. Mol. Biol.* **2017**, *429*, 2556–2570.
- (34) Bosisio, S.; Mey, A. S.; Michel, J. Blinded predictions of distribution coefficients in the SAMPL5 challenge. *J. Comput. Aided Mol. Des.* **2016**, *30*, 1101–1114.
- (35) Mey, A. S.; Juárez-Jiménez, J.; Hennessy, A.; Michel, J. Blinded predictions of binding modes and energies of HSP90- ligands for the 2015 D3R grand challenge. *Bioorg. Med. Chem.* **2016**, *24*, 4890–4899.
- (36) Calabrò, G.; Woods, C. J.; Powlesland, F.; Mey, A. S.; Mulholland, A. J.; Michel, J. Elucidation of Nonadditive Effects in Protein-Ligand Binding Energies: Thrombin as a Case Study. *J. Phys. Chem. B* **2016**, *120*, 5340–5350.
- (37) Mishra, S. K.; Calabró, G.; Loeffler, H. H.; Michel, J.; Koča, J. Evaluation of Selected

- Classical Force Fields for Alchemical Binding Free Energy Calculations of Protein-Carbohydrate Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3333–3345.
- (38) Kuhn, M.; Firth-clark, S.; Tosco, P.; Mey, A. S. J. S.; Mackey, M.; Michel, J. Assessment of Binding Affinity via Alchemical Free-Energy Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 3120–3130.
- (39) Lee, W.-G.; Gallardo-Macias, R.; M. Frey, K.; A. Spasov, K.; Bollini, M.; S. Anderson, K.; L. Jorgensen, W. Picomolar Inhibitors of HIV Reverse Transcriptase Featuring Bicyclic Replacement of a Cyanovinylphenyl Group. *J. Am. Chem. Soc.* **2013**, *135*, 16705–16713.
- (40) Lee, W.-G.; H. Chan, A.; A. Spasov, K.; S. Anderson, K.; L. Jorgensen, W. Design, Conformation, and Crystallography of 2-Naphthyl Phenyl Ethers as Potent Anti-HIV Agents. *ACS Med. Chem. Lett.* **2016**, *7*, 1156–1160.
- (41) Thakur, V. V.; Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. Optimization of pyrimidinyl- and triazinyl-amines as non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem.* **2006**, *16*, 5664–5667.
- (42) Ruiz-Caro, J.; Basavapathruni, A.; Kim, J. T.; Bailey, C. M.; Wang, L.; Anderson, K. S.; Hamilton, A. D.; Jorgensen, W. L. Optimization of diarylamines as non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem.* **2006**, *16*, 668–671.
- (43) Jorgensen, W. L. Computer-aided discovery of anti-HIV agents. *Bioorg. Med. Chem.* **2016**, *24*, 4768–4778.
- (44) Sarafianos, S. G.; Marchand, B.; Das, K.; Himmel, D.; Parniak, M. A.; Hughes, S. H.; Arnold, E. Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J. Mol. Biol.* **2009**, *385*, 693–713.

- (45) Ray, A. S.; Yang, Z.; Shi, J.; Hobbs, A.; Schinazi, R. F.; Chu, C. K.; Anderson, K. S. Insights into the molecular mechanism of inhibition and drug resistance for HIV-1 RT with carbovir triphosphate. *Biochem.* **2002**, *41*, 5150–5162.
- (46) Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. FEP-guided selection of bicyclic heterocycles in lead optimization for non-nucleoside inhibitors of HIV-1 reverse transcriptase. *J. Am. Chem. Soc.* **2006**, *128*, 15372–15373.
- (47) Jorgensen, W. L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K. S.; Hamilton, A. D. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem.* **2006**, *16*, 663–667.
- (48) Schrödinger Release 2020-3: Desmond Molecular Dynamics System. 2020; <https://www.schrodinger.com/desmond>.
- (49) Frisch, M. J. et al. Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.
- (50) Dodda, L. S.; De Vaca, I. C.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: An automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336.
- (51) Chai, J. D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
- (52) Manz, T. A.; Sholl, D. S. Chemically meaningful atomic charges that reproduce the electrostatic potential in periodic and nonperiodic materials. *J. Chem. Theory Comput.* **2010**, *6*, 2455–2468.
- (53) Lee, L. P.; Limas, N. G.; Cole, D. J.; Payne, M. C.; Skylaris, C. K.; Manz, T. A.

- Expanding the scope of density derived electrostatic and chemical charge partitioning to thousands of atoms. *J. Chem. Theory Comput.* **2014**, *10*, 5377–5390.
- (54) Qiu, Y.; Smith, D. G.; Stern, C. D.; Feng, M.; Jang, H.; Wang, L. P. Driving torsion scans with wavefront propagation. *J. Chem. Phys.* **2020**, *152*, 244116.
- (55) Vilseck, J. Z.; Armacost, K. A.; Hayes, R. L.; Goh, G. B.; Brooks, C. L. Predicting Binding Free Energies in a Large Combinatorial Chemical Space Using Multisite λ Dynamics. *J. Phys. Chem. Lett.* **2018**, *9*, 3328–3332.
- (56) Frey, K. M.; Bollini, M.; Mislak, A. C.; Cisneros, J. A.; Gallardo-Macias, R.; Jorgensen, W. L.; Anderson, K. S. Crystal structures of HIV-1 reverse transcriptase with picomolar inhibitors reveal key interactions for drug design. *J. Am. Chem. Soc.* **2012**, *134*, 19501–19503.
- (57) Lim, N. M.; Wang, L.; Abel, R.; Mobley, D. L. Sensitivity in Binding Free Energies Due to Protein Reorganization. *J. Chem. Theory Comput.* **2016**, *12*, 4620–4631.
- (58) Cole, D. J.; Tirado-Rives, J.; Jorgensen, W. L. Enhanced Monte Carlo sampling through replica exchange with solute tempering. *J. Chem. Theory Comput.* **2014**, *10*, 565–571.
- (59) 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.
- (60) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.
- (61) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochow, 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.

- (62) Chen, T.; Manz, T. A. A collection of force field precursors for metal-organic frameworks. *RSC Adv.* **2019**, *9*, 36492–36507.

Graphical TOC Entry

