An Automated On-The-Fly Optimization of Resource Allocation for High-Throughput Protein-Ligand Binding Free Energy Simulations

S. Benjamin Koby, Evgeny Gutkin, Shree Patel, and Maria G Kurnikova

Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, United States

Abstract

Molecular dynamics simulations to compute protein to small molecule binding free energies are becoming a valuable tool in the early stages of drug discovery. However, their cost and complexity are often prohibitive for high-throughput studies. Herein, we present an automated workflow for the thermodynamic integration scheme with the "on-the-fly" optimization of computational resource allocation for each λ-window of both relative and absolute binding free energy simulations. This iterative workflow utilizes automatic equilibration detection and convergence testing via the Jensen-Shannon distance to determine optimal simulation stopping points in an entirely data-driven manner. We benchmark our workflow on the well-characterized systems cyclin-dependent kinase 2 and T4 Lysozyme L99A/M102Q mutant, as well as the more flexible SARS-CoV-2 papain-like protease. We demonstrate that this proposed protocol can achieve over an 85% reduction in computational expense while maintaining similar levels of accuracy when compared to other benchmarking protocols. We examine the performance of this protocol on both small and large molecular transformations. The cost accuracy tradeoff of repeated runs is also investigated.

Introduction

Alchemical binding free energy (BFE) simulations are a class of free energy molecular dynamics (MD) simulations that utilize an "alchemical" i.e., nonphysical thermodynamic pathway to compute free energy differences between environmental conditions or pairs of molecules. In the context of biomolecular systems, MD BFE simulations are often used to compute the binding free energy of a small molecule ligand to a protein,[1,2] which is an important parameter in the initial stages of the drug discovery pipeline. The alchemical approach is useful because it is both rigorous, as the free energy of the system is a state function, and significantly more efficient than brute force

unbiased simulations,[3] though the method suffers from errors related to imperfect forcefields, rough free energy surfaces, and other technical challenges. There are two broad classes of BFE simulations: absolute binding free energy (ABFE) simulations, and relative binding free energy (RBFE) simulations. ABFE simulations seek to calculate the difference in free energy between the protein-ligand complex and the free protein and ligand in solution, referred to as the standard binding free energy ΔG_{bind}. This quantity can be measured experimentally, e.g., by surface plasmon resonance[4] or isothermal titration calorimetry.[5] ΔG_{bind} is related to the dissociation constant K_d by the following relation:

$$\Delta G_{bind} = RT \ln K_d, \tag{1}$$

where R is the gas constant and T is the temperature of the system. RBFE calculations, on the other hand, seek to calculate the difference in ΔG_{bind} of two ligands, referred to as $\Delta\Delta G_{bind}$. In drug design applications ABFE simulations are used for initial screenings and hit identification of compounds for synthesis,[6] while RBFE simulations are well-suited for hit-to-lead and lead optimization.[3,7–10]

Despite their greater computational expense and complexity, MD-based BFE simulations have several advantages over other, simplified computational methods. For example, while molecular docking, a commonly used computational screening technique, can quickly yield reasonable ligand binding poses, typical scoring functions are often not very useful in discriminating false positives from true hits,[6,11–13] which is a serious issue when thousands of potential hits have been proposed with limited resources for experimental validation. End-point MD methods, such as MM/PBSA and MM/GBSA, may sometimes afford limited accuracy improvement when compared to docking, albeit at a greater cost; yet, their accuracy is extremely system specific and is less reliable than alchemical BFE simulations.[14,15]

Alchemical BFE simulations are designed to transform one molecular system into another via performing stratified MD simulations along a pre-defined reaction pathway. Typically the pathway is defined as a liner interpolation between Hamiltonians of two systems A and B:

$$V(\lambda) = \lambda V_A + (1 - \lambda)V_B, \tag{2}$$

2

$V(\lambda) = \lambda V_A + (1 - \lambda)V_B$ where *V(λ) is the coupled potential function, $\lambda \in [0,1]$* defines the reaction coordinate and $V_{A/B}$ is the potential of endpoint *A/B*, respectively. Usually, each MD simulation about the reaction pathway is independent of the others. Some simulation schemes, however, do not employ independent simulations. These include Hamiltonian exchange variants involving the exchange of neighboring strata,[9,16] as well as schemes such as that employed by He *et al.*,[17] where initial structures for simulation were obtained from snapshots of simulations from neighboring strata. The binding free energy of a transformation is obtained by use of an estimator, most often the Bennet acceptance ratio[18] (BAR), multistate BAR[19] (MBAR) or thermodynamic integration[20] (TI) methods, which combine data from the various strata.[21,22] In TI, the estimator employed in this work, the free energy difference between endpoints A and B are derived by integrating the derivative of the coupled potential function with respect to λ. In practice, this is performed numerically:

$$\Delta G_{A \to B} = \int_0^1 \langle \frac{dV}{d\lambda} \rangle_\lambda \, d\lambda \cong \sum_i^N w_i \langle \frac{dV}{d\lambda} \rangle_i, \tag{3}$$

where $\langle \frac{dV}{d\lambda} \rangle_i$ is the average of the derivative of the coupled potential function over the MD simulation at *λ = i*, w_i is the statistical weight of the strata determined by the selected integration scheme, and *N* is the number of strata employed.

With current force fields, software and sampling scheme implementations, RBFE simulations can routinely achieve mean absolute errors (MAE) of approximately 1 kcal/mol when compared to experimental results.[9,23,24] ABFE simulations are more challenging, and while several groups have reported MAE similar to that of RBFE,[13,25,26] a recent meta-analysis by Fu *et al.*[27] found that alchemical ABFE simulations have an average MAE of 1.88 kcal/mol. With the advent of GPU computing, alchemical BFE simulations are completable on a timescale that is relevant for drug discovery applications;[7,17,28–30] however, computing the BFEs of large numbers of ligands remains prohibitively costly and time consuming.[31] Moreover, the technique suffers from various technical difficulties, including the tedious nature of system setup and preparation; inaccuracies in predicting difficult transformations such as ring breaking/forming, ring extensions, and change in net charge;[3,32–35] and the difficulty in achieving sufficient sampling across all intermediate states. Combined, these problems often prove unsurmountable, leading to the selection of cheaper and simpler methods with inferior accuracy.

Recently, we have developed a hit-to-lead optimization framework combining RBFE simulations and active learning (AL) machine learning (ML) to iteratively explore large chemical spaces consisting of thousands of congeneric molecules for high performing molecules when compared to an experimentally validated reference compound.[7] In this scheme, batches of molecules are iteratively selected by an ML model trained on previous iterations of RBFE calculations, with the goal of selecting molecules with the most negative $\Delta\Delta G_{bind}$ value. Other groups have also demonstrated the active learning framework efficacy in RBFE simulations,[36,37] as well as in docking,[38–40] forcefield development[41] and course-graining.[42] Recent work by Thompson *et al.* demonstrated that, using a dataset of 10^4 RBFE simulations performed using a radial perturbation map with a single reference compound, AL was able to select the majority of the molecules with the most negative $\Delta\Delta G_{bind}$ within five cycles of sixty or more molecules per cycle.[43] This method would still entail no fewer than 300 RBFE simulations, which currently requires significant resources to achieve. On the other hand, ABFE simulations have been proposed as an accurate and cost-effective method for the final stages of high-throughput virtual drug screening for initial hit discovery.[6] Beginning with large datasets of molecules, often in the millions to billions, this approach utilizes high-throughput docking to narrow down the set of candidates to a number feasible for ABFE simulations, typically in the hundreds or thousands, with top performing candidates submitted for experimental validation.[6,12] The utility of combining docking with alchemical BFE simulations has been recognized for several decades due to their complementary nature: the inaccuracy of docking can be corrected with ABFE rescoring, while docking both significantly reduces the number of ABFE simulations required and provides reasonable starting poses.[44] Still, this may involve thousands of tedious and expensive calculations, which can be infeasible to perform. Thus, both RBFE and ABFE simulations would benefit from methods designed to lessen the cost of such high-throughput computations.

It is typical to uniformly allocate computational resources by simulating each strata, typically denoted a λ-window, for the same amount of simulation time;[7,9,16,17,23] however, there is no theoretical basis for this practice. For instance, it is unlikely that simulation times necessary for convergence would be equivalent between different systems, such as the protein-ligand complex and the solvated ligand, even if we restrict our analysis to states with identical λ-values. One would expect this likelihood to decline further when considering different ligands or even different protein systems. Furthermore, it is not uncommon for specific λ-windows of a given alchemical

4

transformation to experience larger autocorrelation times than other λ-windows, resulting in slower convergence, fewer uncorrelated samples, and greater statistical uncertainty. In contrast, other λ-windows converge very quickly, with additional sampling affording marginal at best accuracy benefits. In short, using a uniform allocation of resources may result in wasting resources simulating already converged λ-windows, while starving more difficult λ-windows of resources. This issue is magnified when considering a high-throughput computational drug discovery campaign due to the large number of necessary simulations.

Herein, we present a simple, highly automated, and data driven scheme for on-the-fly optimization of computational resource allocation for each λ-window of high-throughput RBFE and ABFE simulations using TI. The goal of this scheme is to provide the minimal resources necessary to achieve a convergence threshold for each λ-window with as little user input as possible. We begin by describing the theory behind alchemical BFE simulations and the simulation parameters employed for each system studied in this work. Next, we illustrate the workflow and concepts behind our on-the-fly optimization algorithm. We then demonstrate our RBFE implementation on the cyclin-dependent kinase 2 (CDK2) benchmark system using the same input files as Song *et al*.[23] and compare our accuracy with respect to experimental results against theirs. In order to examine the performance of our protocol on more difficult and flexible systems, we apply several implementations of our protocol to several ligand mutations of the SARS-CoV-2 papain-like protease (PLpro) that we have performed in previous work.[7] As PLpro is less characterized as CDK2, we compare our results against long simulations as opposed to experimental results. We test our ABFE implementations on the T4 Lysozyme L99A/M102Q mutant in complex with N-phenylglycinonitrile (PDB ID: 2RBN[45]) and compare our calculated binding affinity against the experimental value. We also compare the results of several of our protocol implementations on the PLpro ligand *N*-[(1*R*)-1-naphthalen-1-ylethyl]benzamide against long simulations. We demonstrate that our protocol maintains accuracy while yielding increases in computational efficiency when compared to previously published sampling schemes. Finally, we conclude by making prescriptions for future high-throughput alchemical BFE drug discovery campaigns.

Methods

Theory. The thermodynamic cycle for an RBFE MD simulation is depicted below in Figure 1A. The difference in ΔG_{bind} between two ligands is computed by performing an alchemical mutation of the ligands from one to the other in both the bound and unbound states. As one can see in Figure 1, the difference in ΔG_{bind} of two compounds A and B, denoted $\Delta\Delta G_{bind}^{A\rightarrow B}$, is equivalent to the difference in free energies of the compounds bound to the protein and in solution:

$$\Delta\Delta G_{bind}^{A\rightarrow B} = \Delta G_{bind}^{A} - \Delta G_{bind}^{B} = \Delta G_{A\rightarrow B}^{prot} - \Delta G_{A\rightarrow B}^{wat} \tag{4}$$

Thus, RBFE calculations consist of two alchemical transformations: from one ligand to the other while in complex with the protein, and from one ligand to the other while in solution. While in principle any mutation could be accomplished, only minor mutations are commonly performed due to the requirements that ligands have the same common substructure and that binding modes be preserved across ligands.[7–9,23] Therefore, RBFE campaigns are typically restricted to sets of congeneric ligands that share a common substructure but differ in the identity of appendage functional groups.

The thermodynamic cycle for an ABFE MD simulation is depicted below in Figure 1B. The free energy difference between the protein-ligand complex and the free protein and ligand in solution ($1 \rightarrow 2$: ΔG_{bind}°) is calculated by completing an alchemical pathway between these two endpoint states. A key component of these simulations is the virtual bond, first introduced by Boresch *et al.*,[46] which is a set of harmonic restraints between three protein backbone heavy atoms and three ligand heavy atoms (see Fig. S1). The application of the virtual bond is discussed below. ABFE simulations typically consist of three alchemical transformations: 1) within the protein-ligand complex, the addition of a virtual bond between three protein residue backbone atoms and three ligand heavy atoms ($2 \rightarrow 4$: ΔG_{+VB}^{prot}); 2) within the protein-ligand complex with the virtual bond, the removal of electrostatic and van der Waals interactions either stepwise or simultaneously, often termed "annihilation", of the ligand ($4 \rightarrow 5$: ΔG_{int}^{prot}); and 3) the annihilation of the free ligand in solution ($1 \rightarrow 3$: ΔG_{int}^{solv}). As Boresch *et al.* demonstrated, once the ligand with the virtual bond has been annihilated, the free energy of the virtual bond can be calculated analytically as the ligand is otherwise non-interacting with the environment[46] ($5 \rightarrow 6$: ΔG_{-VB}^{prot});

thus, after its removal, there is no free energy cost to remove the ligand from the binding pocket $(6 \rightarrow 3)$. Thus, we can calculate ΔG_{bind} by the following equation:

$$\Delta G_{bind}^{\circ} = \Delta G_{int}^{solv} - \Delta G_{+VB}^{prot} - \Delta G_{-VB}^{prot^{\circ}} - \Delta G_{+VB}^{prot} - \Delta G_{int}^{prot}, \tag{5}$$
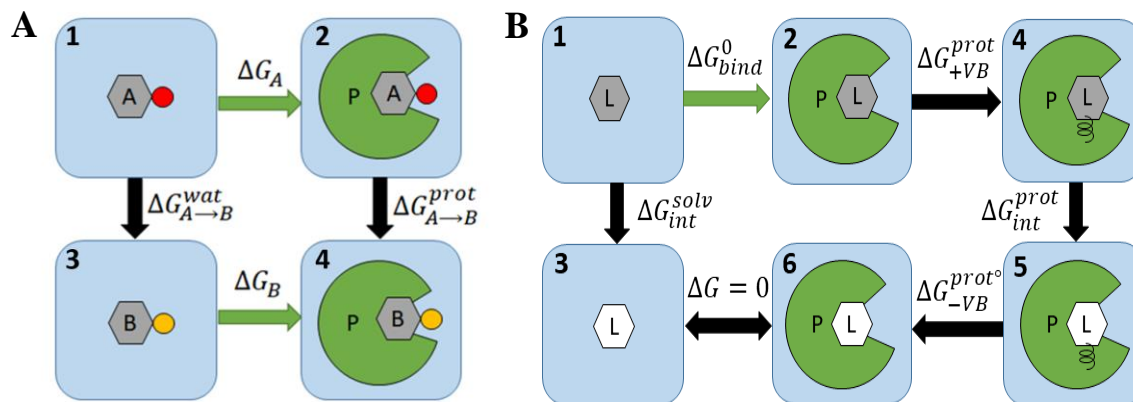

Figure 1. Alchemical Thermodynamic Cycles for RBFE (left) and ABFE (right). The arrows determine the direction of the transformation, and therefore the sign of the term in the sum in Equations 4 and 5.

Software. All MD simulations were performed using the GPU-accelerated pmemd.cuda module of AMBER20.[47–50] All $\frac{dV}{d\lambda}$ gradient timeseries data (see Eq 3) were extracted with the alchemlyb[51] python package. Decorrelation was performed using the pymbar[52] python package, whereby decorrelated samples were obtained by subsampling with the statistical inefficiency rounded up to the nearest integer value. All hydrogen bonding interactions between the protein and ligand were analyzed with CPPTRAJ.[53] All input coordinates, topology and parameters for conventional MD simulations were obtained using Ambertools18.[54]

RBFE CDK2 system setup and TI simulations. Simulation starting structures and input files were extracted from the GitHub repository https://github.com/linfranksong/Input_TI.[23] For both the protein-ligand complex and solvated ligand mutations, two separate λ-schedules were used. The first employed the following twelve λ-windows: 0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, and 0.99078. The second employed the following nine λ-windows: 0.01592, 0.08198, 0.19331, 0.33787, 0.5, 0.66213, 0.80669, 0.91802 and 0.98408. For each λ-window, the following protocol was employed: 1) 2000 steps of minimization with the gradient descent method; 2) 50 ps of heating from 0.1 K to 300 K

7

in the NVT ensemble; 3) 300 ps of density equilibration in NPT; and 4) production simulations in NVT with gradient averages obtained via the bootstrap method. Harmonic RMSD restraints were imposed on heavy atoms of the protein and the ligand during minimization and heating and were gradually removed during density equilibration; no restraints were used during production simulations. Special care was given to the λ=0.98408 window of the nine-point λ-schedule to avoid numerical instability. For this window, the structure obtained after the first 1000 steps of minimization with gradient descent of the λ=0.91802 window was used as the input for the second 1000 steps of minimization, and then the protocol proceeded as normal. For the 12-window λ-schedule production simulations, a 1 fs timestep was used without SHAKE and a Berendsen thermostat was employed. This was done to mimic the production protocol of Song *et al.* more closely. For the 9-point λ-schedule, a 2 fs timestep was used with SHAKE and a Langevin thermostat was employed. Free energies of both the protein-ligand complex and solvated ligand alchemical steps were obtained using the gaussian quadrature rule. For each mutation, a total of 10 independent simulations were performed per protocol.

RBFE PLpro system setup and TI simulations. Four ligands were selected from our previous work on the PLpro system and are displayed below in Figure 2. Ligands 1-3 were selected such that their $\Delta\Delta G_{bind}$ calculated in previous work were negative (Ligand 1), approximately 0 (Ligand 2), and positive (Ligand 3). Ligand 4 was selected as an edge case due to the difficulty of the mutation (see Discussion). Input coordinates, topologies, and parameters were obtained from our previous work (see Gusev *et al.*[7] for details). All λ-windows were minimized and equilibrated using the protocol described in the previous section, with restraints applied to two water molecules in the binding pocket during minimization and heating, and gradually removed during density equilibration.
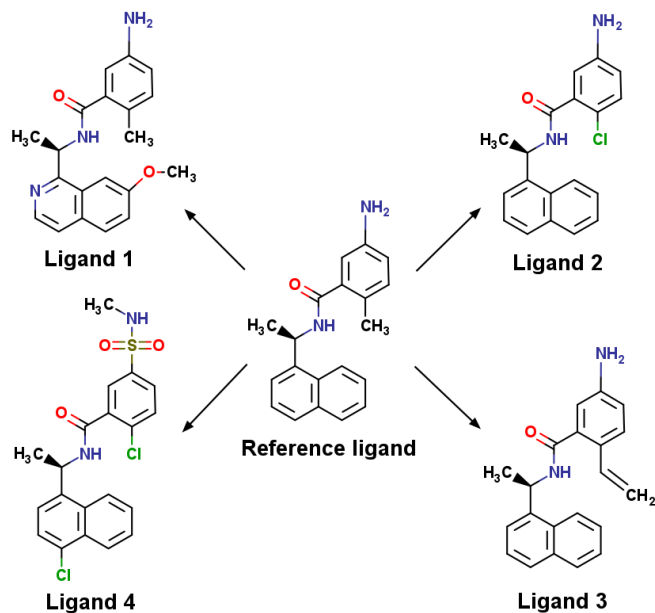
8

Figure 2. PLpro ligands used for ABFE and RBFE simulations. ABFE simulations were performed on the Reference Ligand, while RBFE simulations were performed by mutating the Reference Ligand to Ligands 1, 2, 3 or 4.

Lysozyme protein preparation and simulation. The crystal structure of the T4 lysozyme L99A/M102Q in complex with N-phenylglycinonitrile was extracted from Protein Data Bank (PDB ID: 2RBN[45]). Ligand atom parameters were obtained using GAFF2 (version 2.11)[55] and ligand atomic charges were derived using the RESP[56] method with Gaussian 09[57] or 16.[58] The protein was solvated in an orthorombic TIP3P water box using tleap with 15 Å distance between the protein and the edge of the box. The simulation protocol included the following steps: 1) 2000 steps of minimization with gradient descent method; 2) 100 ps of heating from 1 K to 298 K in NVT ensemble; 3) 300 ps of density equilibration in NPT ensemble; 4) 7 ns of production simulation in NVT. All the MD simulations were performed using a 2 fs time step. Harmonic RMSD restraints were imposed on heavy atoms of the protein and the ligand during minimization and heating and were gradually removed during density equilibration; no restraints were used during production simulations.

Selection of virtual bond atoms. The first 2 ns of the MD production simulation were discarded, and the average structure was obtained from the last 5 ns of the simulation. A trajectory frame with

9

minimum RMSD of ligand heavy atoms with respect to average structure was selected as a representative structure and used as the initial protein-ligand complex structure for TI simulations.

Selection of atoms for the virtual bond approach[46] was performed using a modified version of the algorithm described in Chen *et al.,*[13] and was proceeded as follows: All hydrogen bonding interactions present in 50% or more of the frames were selected and ranked based on the ligand atom distance from the ligand centroid. The atom closest to the centroid was then selected. If the ligand atom was a hydrogen, then the heavy atom bonded to this hydrogen was used instead. All neighboring heavy atoms and their corresponding neighboring heavy atoms were then selected. Protein and ligand atoms were scored based on their angles within the reference frame. Given protein backbone atoms A, B, and C, and ligand atoms a, b, and c, the following angles were calculated: ABC, aAB, baA, and abc (see Fig. S1). Atom combinations in which both angles were between 45° and 135° were considered valid. The possible ligand atom combinations fell into three categories. Within the first category, atom a was a heavy atom involved in hydrogen bonding, atom b a neighboring heavy atom of atom a, and atom c a neighboring heavy atom of atom b. Within the second category, atom a was the heavy atom involved in hydrogen bonding, atom b was a neighboring heavy atom of a neighboring atom of atom a, and atom c was a second neighboring heavy atom of the same neighboring heavy atom of atom a. In the third category, atoms a and c of category two are swapped with each other. For each of these categories, two combinations of protein heavy atoms of residue involved in hydrogen bonding were attempted: N-C_α-C and C- C_α-N. All valid configurations were then scored based on their mean absolute deviation from 90°, and the combination with the lowest score was selected for the virtual bond. If no valid atom combination was found, the next closest ligand atom to the centroid involved in hydrogen bonding was selected and analyzed for valid combinations. This process was repeated until a valid combination was found. If no atom combination was found to be valid, all ligand heavy atoms were ranked based on their distance to the centroid of the ligand. The atom closest to the centroid was selected and compared against every combination of neighboring C_α atoms in the manner described above. If a valid combination was found, then the distance between ligand atom a and protein atom A was calculated. If the distance was found to be less than 1 nm, the combination was accepted as valid. If not, the next closest atom to the centroid was selected, and this process was repeated until a valid combination was found.

ABFE lysozyme TI simulations. For TI simulations of solvated ligand, the ligand was solvated in a TIP3P water box using tleap[54] with a 15 Å distance between the ligand and the edge of the box. For TI simulations of the protein-ligand complex, the orientation of ligand with respect to the protein were restrained using the virtual bond approach.[46] Force constants of 4 kcal/(mol*Å^2), 20 kcal/(mol*rad^2) and 40 kcal/(mol*rad^2) were used for distance, angle and dihedral angle restraints, respectively. The second-order smoothstep softcore potential (SSC(2)), as implemented in AMBER20, was utilized for both the protein-ligand complex and solvated ligand steps. For each λ-window, the system was minimized and then equilibrated using the same protocol as was performed for conventional MD. For the solvated ligand and protein-ligand complex systems, a λ-schedule of 9 equally distributed windows was used (0.1, 0.2, 0.3, …, 0.9). Gradient means and variances were calculated using the bootstrap method.[59] For the addition of the virtual bond restraints, 7 unequally distributed λ-windows were used, with window density concentrated near the minimum λ-windows. Free energies for the ligand, the protein, and the restraint addition were obtained via the trapezoid rule. The free energy of adding virtual bond restraints for the non-interacting ligand was calculated using the Boresch formula.[46] A total of 39 independent simulations were performed for all protocols.

ABFE PLpro system setup and TI simulations. ABFE simulations were performed on *N*-[(1*R*)-1-naphthalen-1-ylethyl]benzamide, henceforth referred to as the "Reference Ligand." Input coordinates, topologies, and parameters were obtained from our previous work (see Gusev *et al.* for details). Force constants of 10 kcal/(mol*Å^2), 10 kcal/(mol*rad^2) and 20 kcal/(mol*rad^2) were used for distance, angle and dihedral angle restraints, respectively. All λ-windows were minimized and equilibrated using the protocol described above, with the addition of harmonic restraints applied to two water molecules in the binding pocket during minimization and heating, and gradually removed during equilibration. No restraints were applied during production.

Approach, Results and Discussion

On-the-fly optimization. Our algorithm for on-the-fly optimization of computational resources used by TI MD simulations proceeded for each λ-window in the following manner (Fig. 3): Starting from the equilibrated structure, an initial short production simulation was performed and $\frac{dV}{d\lambda}$ gradient timeseries (see Eq. 3) were extracted. Convergence testing to determine whether the production simulation should be extended was performed as follows. The gradient timeseries was

11

equilibrated with automatic equilibration detection[60] (AED), as implemented in the pymbar python package. This method determines the optimal equilibration time to be that which maximizes an uncorrelated sample size that can be obtained from an equilibrated gradient timeseries. After equilibration the gradient timeseries was decorrelated (see Methods); then split in half chronologically and each half binned into seven equally spaced bins. The Jensen-Shannon (JS) distance[61], a measure of distance between two probability distributions, between these two histograms was calculated. Given two probability distributions P and Q, the JS distance is defined as follows:

$$JS(P||Q) = \sqrt{\frac{1}{2}(D(P||M) + D(Q||M))}, \tag{6}$$

where $D(P||M)$ is the Kullback-Liebler divergence and $M = \frac{1}{2}(P + Q)$. If the JS distance was greater than or equal to 0.1 or the total number of decorrelated samples was less than 50, another short additional production simulation was performed starting from the last frame of the production trajectory. This was repeated until convergence, or until either of the following two scenarios were met: a total simulation time of 6.5 ns had been achieved and more than 50 decorrelated samples were acquired in total, or a total simulation time of 10.5 ns had been achieved (Fig. 3).
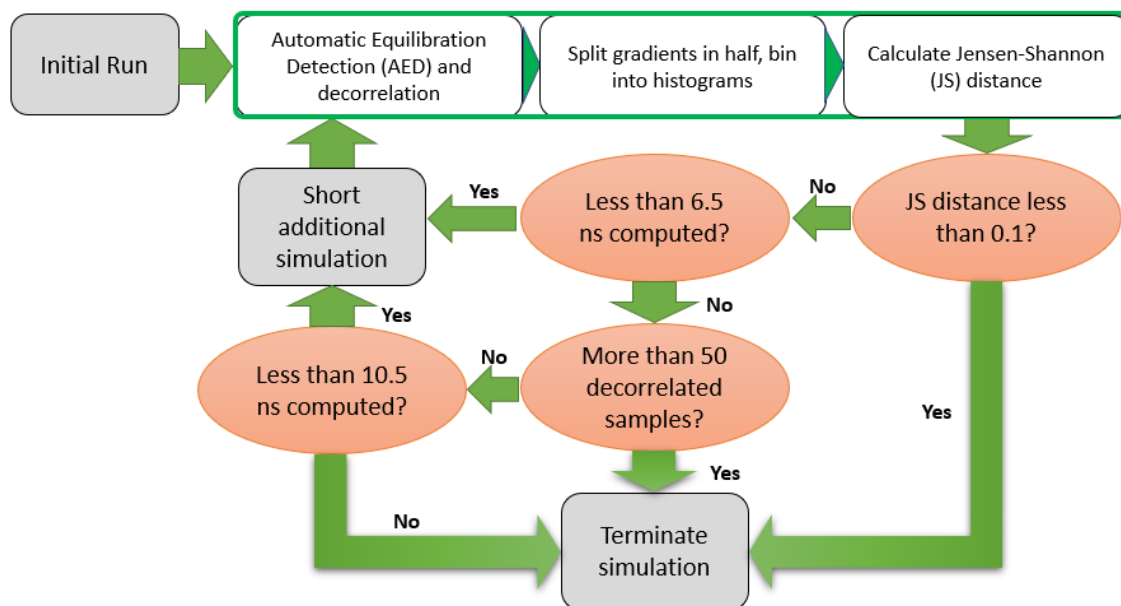

Figure 3. Flowchart of the on-the-fly resource optimization for high-throughput binding free energy TI MD simulations.

12

In this work, we tested six different implementations of this algorithm: Protocols A-E, and C12.

Table 1 Parameters of the various on-the-fly optimization protocols employed.

Protocol	Initial Simulation Length (ns)	Additional Simulation Length (ns)	Number of λ-Windows
A	2.5	0.5	9
B	1.5	0.5	9
C	1.0	0.25	9
D	0.5	0.25	9
E	3.5	0.5	9
F	1.0	0.25	12

Protocols A-E differ solely in the amount of simulation time used in the initial and additional simulation steps. Protocol A uses a 2.5 ns initial simulation with 0.5 ns additional simulations, Protocol B uses a 1.5 ns initial simulation with 0.5 ns additional simulations, Protocol C uses a 1.0 ns initial simulation with 0.25 ns additional simulations, Protocol D uses a 0.5 ns initial simulation with 0.25 ns additional simulations, and Protocol E uses a 3.5 ns initial simulation with 0.5 ns additional simulations. Protocol C12 was only employed for RBFE calculations for CDK2 and utilized the 12-window λ-schedule with a 1.0 ns initial simulation and 0.25 ns additional simulations.

RBFE calculations for CDK2. Experimental $\Delta\Delta G_{bind}$ values for each mutation were obtained from Song *et al.*[23] Overall mean absolute error (MAE) and root mean squared error (RMSE) between experimental and computed $\Delta\Delta G_{bind}$, when considering all 10 replicates per mutation, was 1.03 kcal/mol and 1.24 kcal/mol for Protocol A, 1.02 kcal/mol and 1.29 kcal/mol for Protocol B, 0.99 and 1.25 kcal/mol for Protocol C, and 0.96 and 1.24 kcal/mol for Protocol C12 (see Table 1 for protocol details). The overall R^2 of Protocols A, B, C, and C12 was 0.26, .23, .28, and .30, respectively. A permutation test was performed, whereby 10,000 MAEs and RMSEs were computed by randomly permuting the experimental $\Delta\Delta G_{bind}$ values. For all protocols and both MAE and RMSE, the calculated values were less than 99% of these generated values, which indicates that there is a significant association between our predicted $\Delta\Delta Gs$ and their experimental

13

values. This is important as the null model ($\Delta\Delta G_{bind} = 0$ for each mutation) performs extremely well on this system, as others have noted,[17,23] with MAE and RMSE of 0.95 and 1.23 kcal/mol, respectively. Figure 4 displays simulated distributions of MAEs, RMSEs, and R^2s of Protocol A with respect to the number of replicates, referred to hereafter as batch size, included in the calculation. These distributions were generated by taking 10,000 random combinations of the given sample size of replicates, averaging the $\Delta\Delta G_{bind}$, and then calculating the MAE, RMSE, and R^2, respectively. Tabulated summary statistics for these distributions are shown in Table S1.
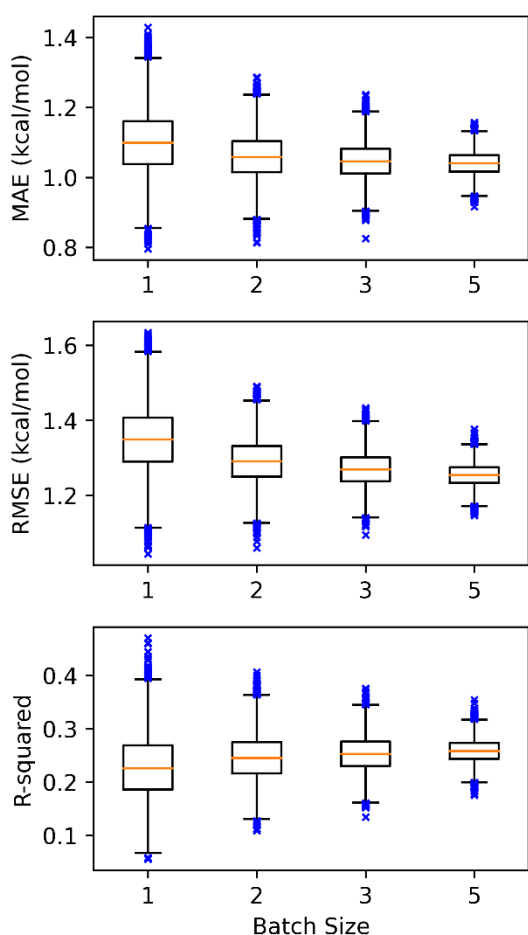

Figure 4. RMSE, MAE, and R^2 of RBFE calculations for CDK2 employing a 9-point Gaussian quadrature λ-schedule using Protocol A with respect to experimental values. Batch size refers to the number of replicates averaged together per mutation.

$\Delta\Delta G_{bind}$ obtained from Protocol C12 and Protocols A, B, and C displayed good agreement, with R^2 values of 0.926, 0.945, and 0.938, respectively, when considering all 10 replicates per mutation, as seen below in Figure 5.
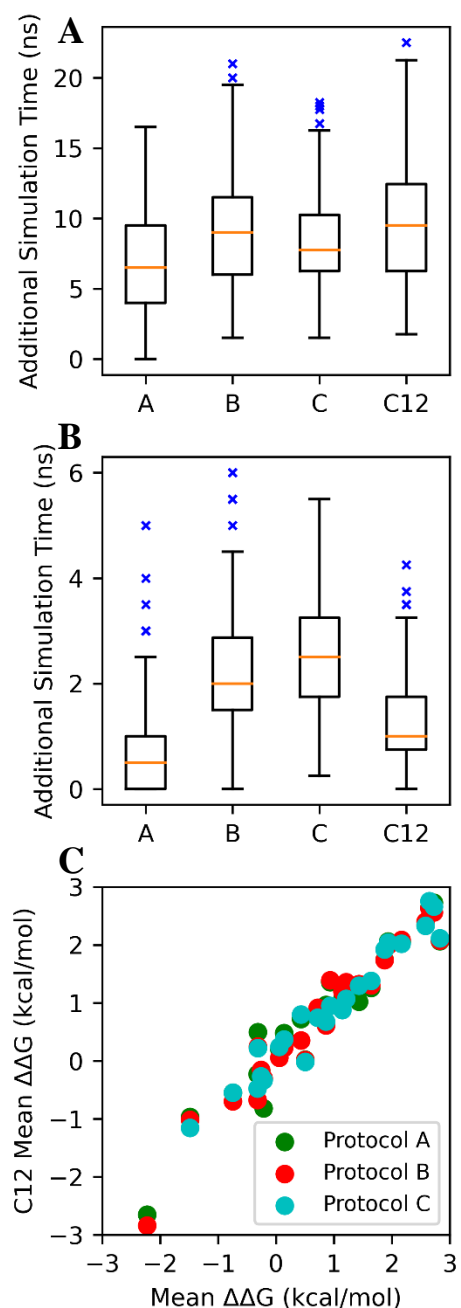

Figure 5. A) Distributions of protein-ligand complex simulation step additional simulation times for all mutations by protocol. B) Distributions of solvated ligand simulation step additional simulation times for

all mutations by protocol. C) Scatterplot of $\Delta\Delta G_{bind}$ values from Protocol C12 versus Protocols A, B, and C.

The number of additional simulations performed varied by protocol, mutation, λ-window, and alchemical step, as seen in Figure 5. As expected, the protein-ligand complex step required more average additional simulation time to achieve convergence than the solvated ligand step regardless of protocol employed. Protocol A generally required the least additional simulation time to achieve convergence in both alchemical steps (see Table 1 for protocol details). Overall, the additional simulation time was evenly distributed between λ-windows of the solvated ligand step, with more variation in the protein-ligand complex step; however, this pattern broke down when comparing specific mutations, as seen in Figure 6. Correlation was observed between the average additional simulation times of the λ-windows of the various protocols, which indicates that our protocol is correctly identifying λ-windows that require additional simulation to achieve convergence.
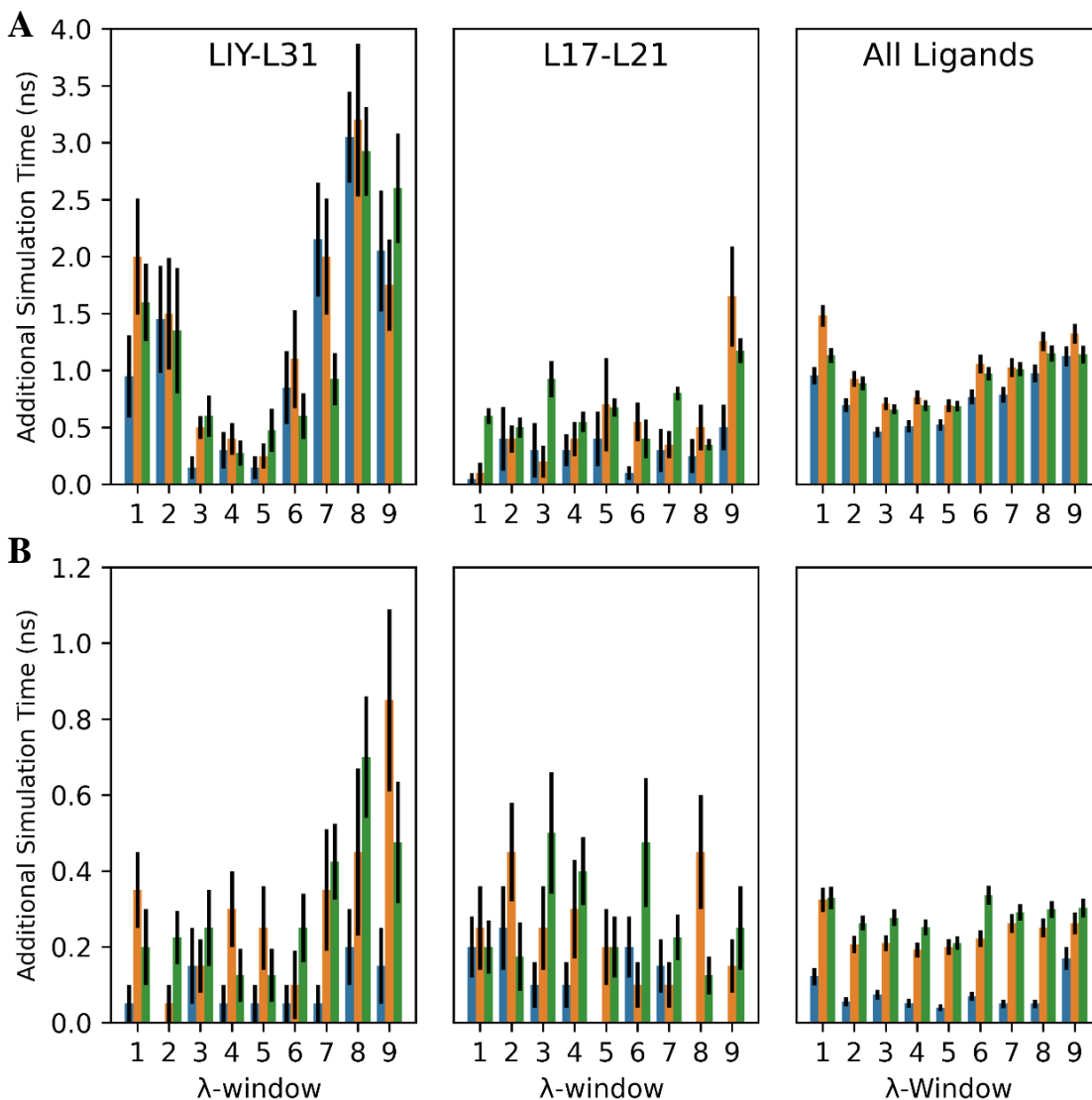
16

Figure 6. A) Distributions of average protein-ligand complex step additional simulation time performed per RBFE simulation of the LIY-L31 and L17-L21 mutations and of all 25 mutations. B) Distributions of average solvated ligand step additional simulation time performed per RBFE simulation of the LIY-L31 and L17-L21 mutations and of all 25 mutations. Protocols A, B, and C are shown in blue, orange, and green, respectively. Error bars represent one standard error of the mean.

Truncated trajectories of all replicates were examined, with 10,000 MAEs and RMSEs generated via the resampling scheme described above. Mean MAE and RMSE and their respective standard errors are displayed below in Figure 7. In all protocols, significant accuracy improvement is observed during the first nanosecond of simulation time, with only minor accuracy improvement observed afterwards. After the initial simulation period, Protocols A and C12 do not display any

17

gain in accuracy, whereas Protocols B and C do (see Table 1 for protocol details). These results suggest that a shorter maximum simulation time may be employed without a loss of accuracy.
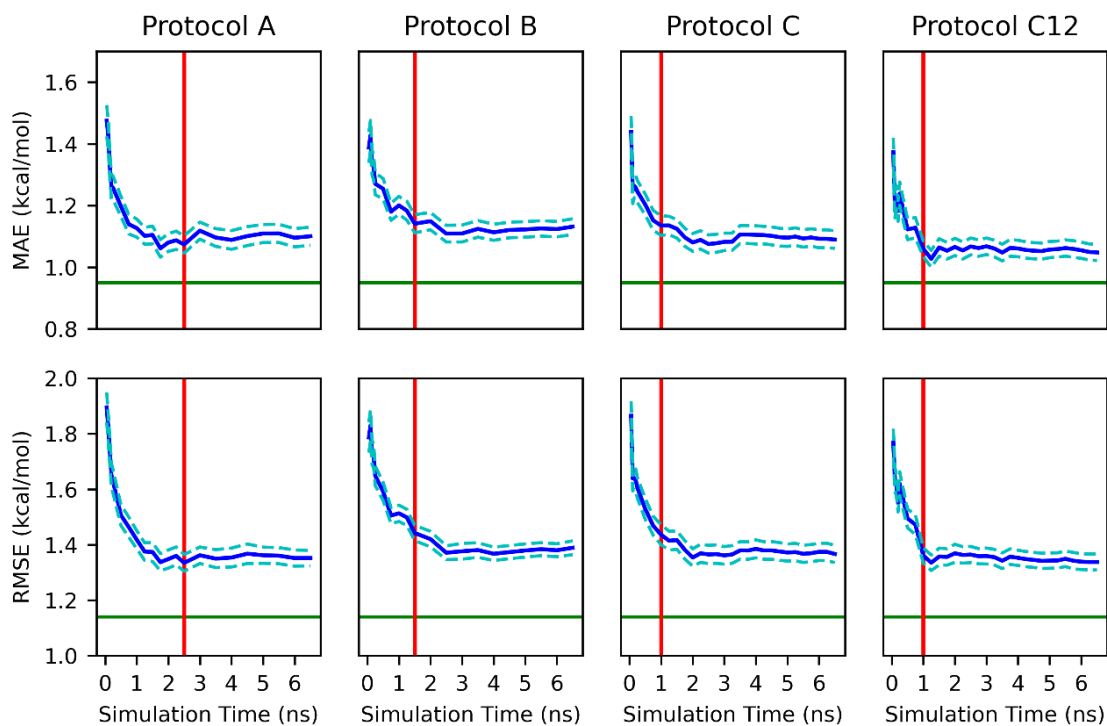

Figure 7. CDK2 MAEs and RMSEs calculated from truncated gradients. The blue line represents the respective loss function (MAE or RMSE), the dashed line represents one standard error of the loss function, the red line represents the length of the initial simulation period, and the green line represents the value of the respective loss function achieved by Song *et al.*

Computational cost savings for both the protein-ligand complex and solvated ligand steps were calculated by taking 10,000 independent samples of a particular batch size of each mutation and summing up the total number of production nanoseconds of a batch of a given mutation for a given alchemical step. For Protocols A, B, and C, the complement of this number divided by 120 was taken, representing the total number of nanoseconds simulated by Song *et al.* for a given mutation multiplied by two to account for their use of 1 fs time step. For Protocol C12, the complement of the sum simulation time divided by 60 was used. Mean and standard errors of savings were calculated from these distributions. Summary statistics of these savings are shown in Table S1.

Overall, the accuracy of the optimized RBFE calculations were comparable to the benchmark results, regardless of protocol employed. The most directly comparable result—a single simulation per mutation using Protocol C12—had an MAE of 1.05±0.08, with comparable results achieved with Protocols A, B, and C (See Table S1). These values are on average slightly greater than the MAE and RMSE of 0.95 and 1.14 kcal/mol, respectively, reported by Song *et al.*, but the difference in MAE is statistically insignificant for Protocols A, C, and C12 at the 95% confidence level. However, these results were achieved with over a 64% reduction in production computational cost. In the case of Protocol C, the average savings were over 85%. All protocols and all batch sizes achieved R^2 values greater than the 0.15 reported by Song *et al.* (see Table S1). Repeated runs, on average, appear to have little benefit, as the average MAE and RMSE decrease only moderately, and the average R^2 increases only slightly, when increasing the batch size from 1 to 10 replicates for all protocols. However, one can see in Figure 4 that the spread of the MAE, RMSE, and R^2 distributions significantly decrease with increasing batch size, meaning the likelihood of calculating a set of particularly poor (or outstanding) RBFEs decreases with repeated runs. This tradeoff should be considered when planning a high-throughput RBFE campaign: if one can accept a higher variance of calculations, then significantly more mutations can be explored at a similar cost.

RBFE calculations for PLpro. For each mutation, three 10-15 ns production simulations were performed at each λ-window (see Methods). The gradients were extracted and analyzed with three different equilibration methods: AED (see On-the-fly optimization), a 2 ns equilibration period, and a 5 ns equilibration period. The equilibrated gradients were then decorrelated and averages were extracted. 10 short-run simulations were performed with Protocols A, B, C, and D for each mutation (see Table 1 for protocol details).

As with the CDK2 mutations, the amount of simulation time applied to each λ-window varied by λ-window, simulation protocol, alchemical step, and mutation, as seen below in Figure 8. We note that λ-windows of protein-ligand complex step with elevated simulation times in the Protocol A also tended to have elevated simulation times in Protocols B-D, and the same pattern holds with respect to Protocols B and C. This demonstrates that different λ-windows converge at different rates and that this pattern is consistent. This pattern is not clear in the solvated ligand step as these λ-windows converged much faster than the protein-ligand complex step.
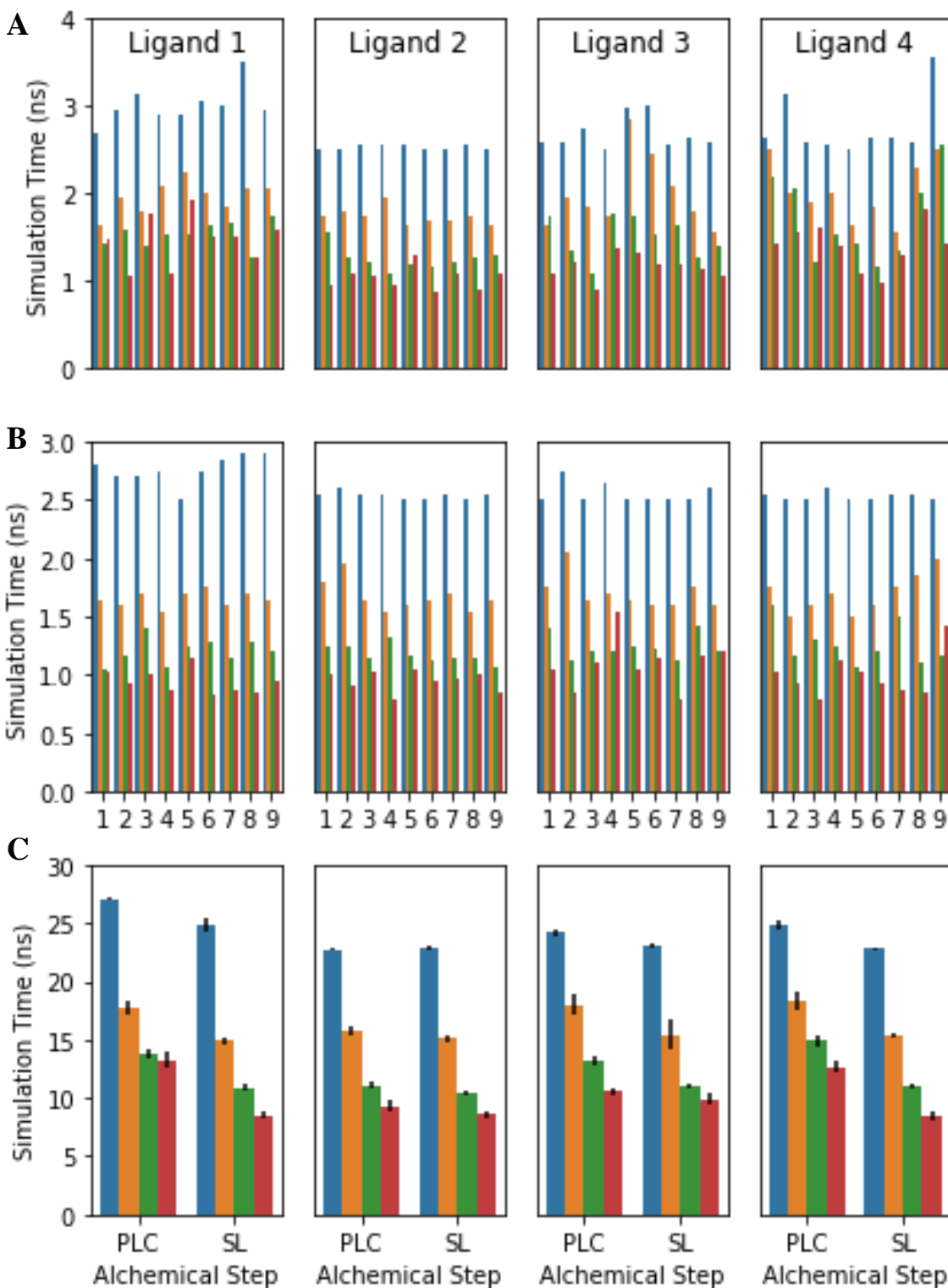
19

Figure 8. A) Average simulation time applied to each λ-window of protein-ligand complex step by protocol and mutation. B) Average simulation time applied to the solvated ligand step by protocol and mutation. C) Overall average applied simulation time per replicate by mutation, alchemical step, and protocol. Blue bars

represent Protocol A, orange bars Protocol B, green bars Protocol C, and red bars Protocol D. Error bars denote one standard error.
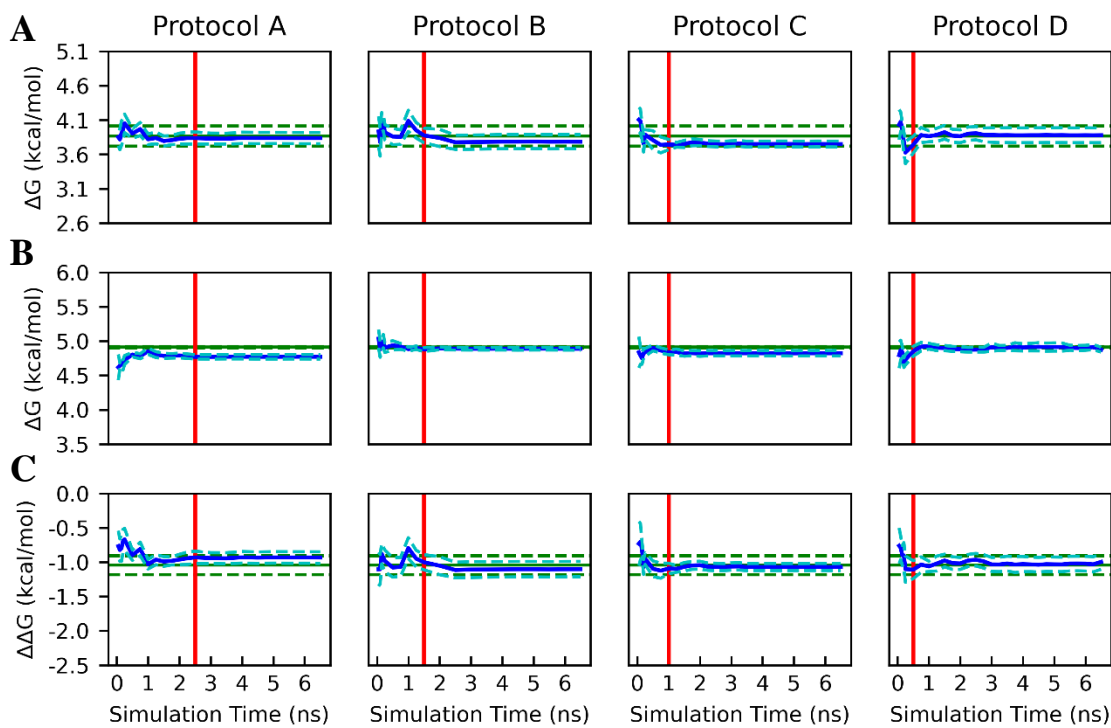

Figure 9. A) Average Ligand 1 protein-ligand complex ΔG calculated from truncated gradients. B) Average Ligand 1 solvated ligand ΔG calculated from truncated gradients. C) Average Ligand 1 $\Delta\Delta G_{bind}$ calculated from truncated trajectories. The blue line represents the mean ΔG or $\Delta\Delta G_{bind}$ and the blue dashed line represents one standard error of the mean. The green line represents the mean long-run ΔG or $\Delta\Delta G_{bind}$ calculated using AED, the green dashed line represents one standard error of the mean, and the red line denotes the length of the initial simulation. Convergence with the long-run simulations was typically achieved quickly, with a similar pattern observed for Ligands 2 and 3 (see Fig. S3 and S4).
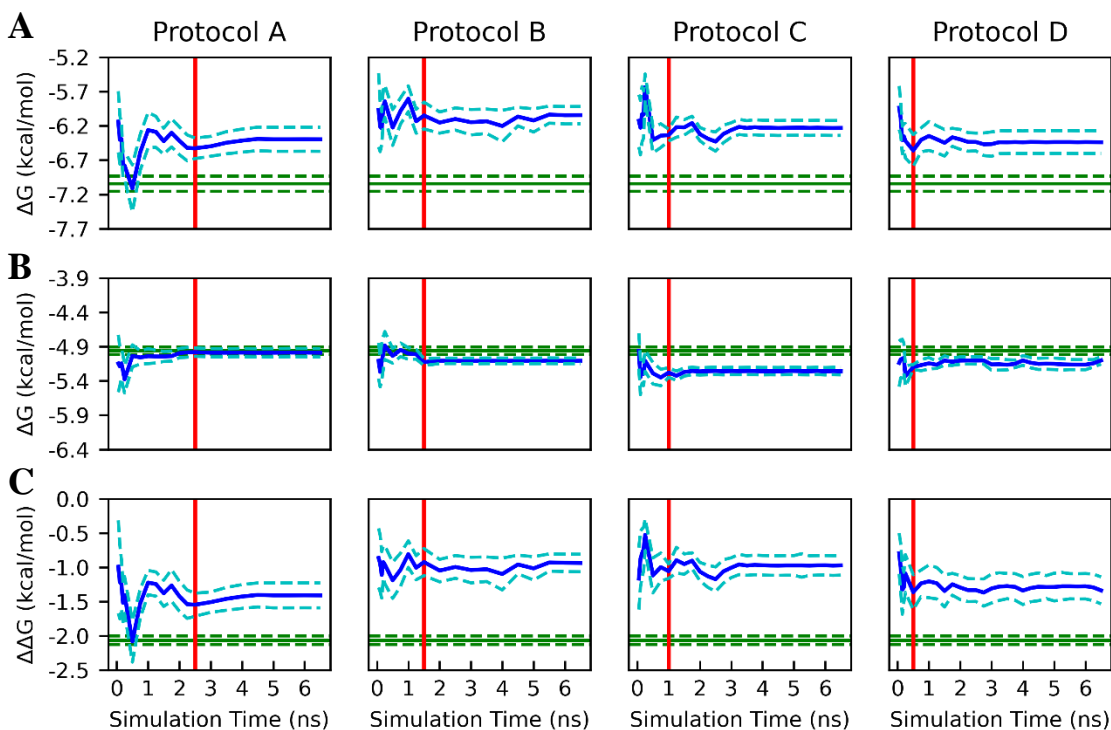
Figure 10. A) Average Ligand 4 protein-ligand complex ΔGs calculated from truncated gradients. B) Average Ligand 4 solvated ligand ΔGs calculated from truncated gradients. C) Average Ligand 4 $\Delta\Delta G_{bind}$ calculated from truncated trajectories. The blue line represents the mean ΔG or $\Delta\Delta G_{bind}$ and the blue dashed line represents one standard error of the mean. The green line represents the mean long-run ΔG or $\Delta\Delta G_{bind}$ calculated using AED, the green dashed line represents one standard error of the mean, and the red line denotes the length of the initial simulation. Significant deviation in the protein-ligand complex step for this difficult mutation is observed for all short-run protocols and the long-run protocols.

Average protein-ligand complex ΔG, solvated ligand ΔG, and overall $\Delta\Delta G_{bind}$ of the short-run simulations and long-run simulations with all three equilibration methods are tabulated in Table 2. $\Delta\Delta G_{bind}$ and alchemical step ΔG values obtained by each short-run protocol were within 1.1 kcal/mol to those obtained by the long-run protocols. Ligand 4 $\Delta\Delta G_{bind}$ values deviated more from their long-run counterparts than those obtained for Ligands 1-3, with absolute error of 0.6-1.1 kcal/mol compared to 0.1-0.2 kcal/mol, respectively. For Ligand 4, the protein-ligand complex step was the major contributor to the absolute error (0.6-1.0 kcal/mol), while the contribution of solvated ligand step was considerably smaller (0.1-0.3 kcal/mol). This deviation did not alleviate with more simulation time in Protocols A-D (Fig. 10), indicating that the relevant timescales are outside the length of these short simulations (see Table 1 for protocol details). The larger error

22

obtained for Ligand 4 can be explained by the difficulty of the mutation, which involves the mutation of the amine group into the methylaminosulfonyl group, as well as a benzene ring methyl group into a chlorine and a naphthalene ring hydrogen into a chlorine. This involves the mutation of seven heavy atoms as well as the overall addition of two rotatable bonds (Fig. 11). The methylaminosulfonyl group, which is flexible and solvent exposed, covers a larger area of the phase space and is thus more difficult to converge and requires more simulation time, as seen in Figure 12. Due to the presence of two rotatable bonds (C-S and S-N), the conformation of methylaminosulfonyl group can vary significantly during RBFE simulations (see Fig. 11 D-E and S5-S6). These conformations differ by interactions with the closest protein residues (D164, E167, Y268 and Q269; see Fig. 11 B-C). This results in considerable fluctuations in $\frac{dV}{d\lambda}$ gradient timeseries and therefore in a slow convergency of $\langle\frac{dV}{d\lambda}\rangle$ at some λ-windows (see Fig S7).
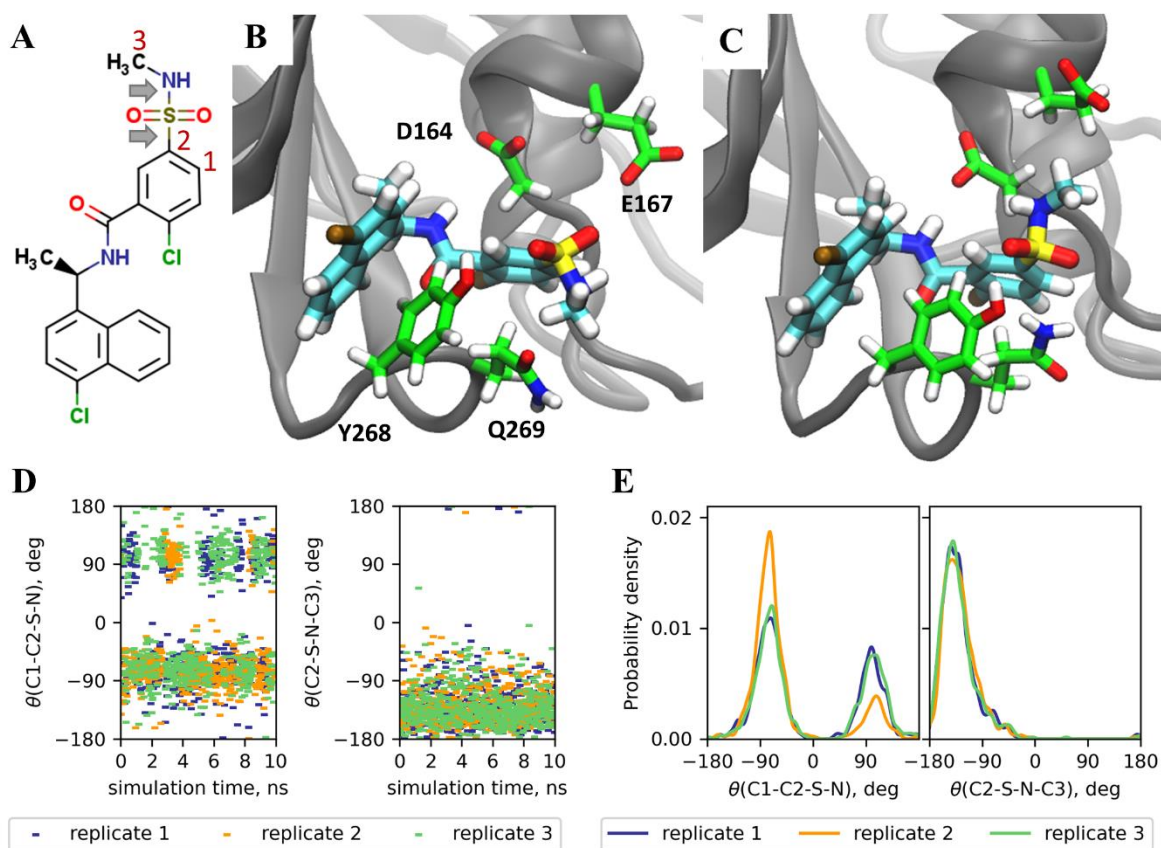

Figure 11. Fluctuations of the methylaminosulfonyl group of Ligand 4 in the protein-ligand complex step of RBFE simulations for PLpro. A) Structure of Ligand 4. Rotatable bonds of the methylaminosulfonyl group are indicated by grey arrows. Carbons atoms forming the corresponding dihedral angles are

23

numbered. B-C) The structures of PLpro in complex with Ligand 4 for the ligand conformation with C1-C2-S-N dihedral angle of approximately 90° (B) and -90° (C) extracted from long-run MD simulations at $\lambda = 0.5$. The protein backbone is shown as the grey cartoon, the ligand and sidechains of residues within 5 Å of the methylaminosulfonyl group are shown as sticks. D) The fluctuation of dihedral angles C1-C2-S-N (left) and C2-S-N-C3 (right) during long-run MD simulations at $\lambda = 0.5$. The three independent replicates of long-run MD simulations are shown by green, orange and blue. E) Distribution of dihedral angles C1-C2-S-N (left) and C2-S-N-C3 (right) at long-run MD simulations at $\lambda = 0.5$.
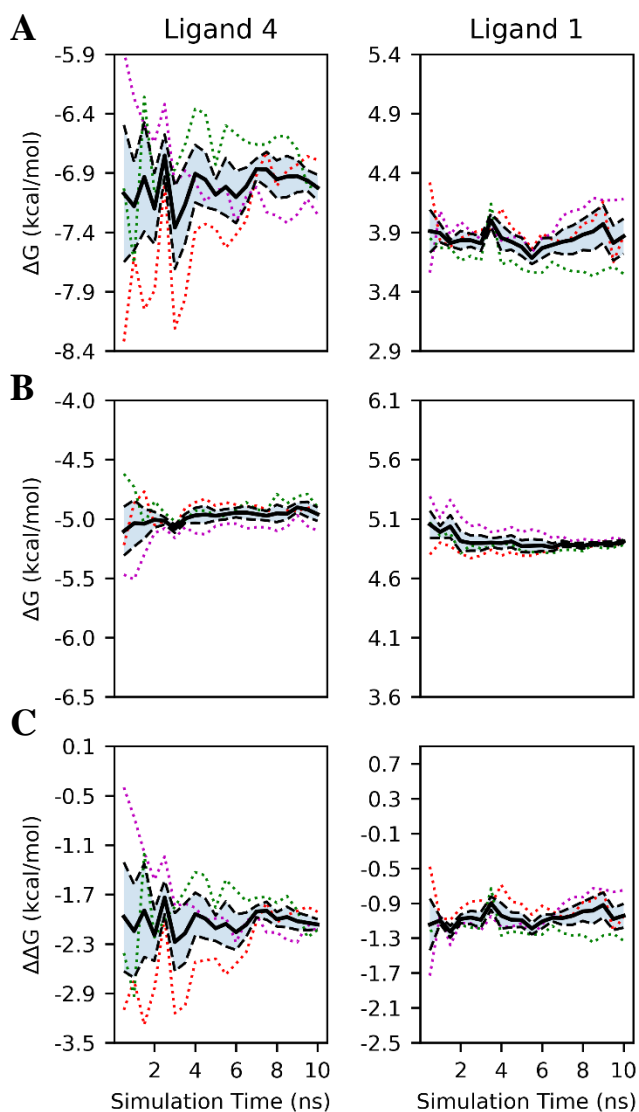

Figure 12. A) Protein-ligand complex ΔGs from truncated gradients of long-run simulations calculated with AED. B) Solvated ligand ΔGs from truncated gradients of long-run simulations calculated with AED. C) $\Delta\Delta G_{bind}$ values from truncated gradients of long-run simulations calculated with AED. Dotted lines

24

represent individual replicates while the black solid and dashed lines represent means and standard errors, respectively. Ligand 4 does not converge until approximately 4 ns of simulation time, whereas Ligand 1 converges much more quickly.

To evaluate this discrepancy, we tested Protocol E on Ligand 4 (see Table 1 for protocol details), which yielded mean values of -6.61 ± 0.12, -5.01 ± 0.04, and -1.60 ± 0.12 kcal/mol for the protein-ligand complex step, the solvated ligand step, and the overall calculation, respectively. Protocol E therefore provided the smallest absolute error in $\Delta\Delta G_{bind}$ with respect to long runs (0.46-0.49 kcal/mol) compared to other protocols. This suggests that for difficult mutations, longer initial simulation times are necessary.

For the protein-ligand complex step of Ligand 4, Protocols A and D obtained the smallest deviation (~0.6 kcal/mol) regardless of equilibration protocol (see Table 1 for protocol details). For Ligands 1-3, all protocols performed similarly. As one can see from Figure 9, convergence with the long-run simulations was achieved quickly, which supports the use of short protocols for these smaller RBFE mutations. Overall, Protocols B utilized approximately 75% and 65% of the computational resources utilized by Protocol A for the protein-ligand complex step and solvated ligand step, respectively, whereas Protocols C and D both utilized approximately 50% and 45%, respectively. This was achieved with no accuracy penalty for Ligands 1-3 and a moderate accuracy penalty for Ligand 4 (~0.47, ~0.43, and ~0.07 kcal/mol, respectively).

Table 2. Average RBFEs and their components of short-run and long-run simulations for PLpro by alchemical step, ligand (L), equilibration method (eq.), and protocol.

Alchemical step	L	Long run (kcal/mol)			Short run (kcal/mol)			
		AED eq.	2 ns eq.	5 ns eq.	Protocol A	Protocol B	Protocol C	Protocol D
Protein-Ligand Complex	1	3.87 ± 0.15	3.80 ± 0.11	3.80 ± 0.16	3.84 ± 0.08	3.78 ± 0.10	3.75 ± 0.04	3.88 ± 0.10
	2	-1.96 ± 0.03	-1.98 ± 0.03	-1.99 ± 0.04	-1.93 ± 0.02	-1.90 ± 0.03	-1.98 ± 0.05	-1.79 ± 0.03
	3	4.75 ± 0.11	4.82 ± 0.09	4.82 ± 0.14	4.80 ± 0.04	4.83 ± 0.04	4.75 ± 0.04	4.90 ± 0.07
	4	-7.02 ± 0.13	-7.03 ± 0.22	-6.97 ± 0.28	-6.40 ± 0.18	-6.04 ± 0.13	-6.23 ± 0.12	-6.44 ± 0.17
Solvated Ligand	1	4.91 ± 0.01	4.84 ± 0.02	4.88 ± 0.04	4.77 ± 0.03	4.89 ± 0.02	4.82 ± 0.04	4.87 ± 0.03
	2	-2.06 ± 0.01	-2.07 ± 0.01	-2.06 ± 0.05	-2.09 ± 0.02	-2.10 ± 0.00	-2.12 ± 0.02	-1.98 ± 0.03
	3	3.50 ± 0.02	3.50 ± 0.02	3.53 ± 0.01	3.52 ± 0.02	3.54 ± 0.00	3.51 ± 0.02	3.53 ± 0.03
	4	-4.96 ± 0.07	-4.94 ± 0.06	-4.90 ± 0.10	-4.99 ± 0.06	-5.11 ± 0.01	-5.26 ± 0.06	-5.10 ± 0.05
Total RBFE ($\Delta\Delta G_{bind}$)	**1**	-1.04 ± 0.14	-1.04 ± 0.09	-1.09 ± 0.13	-0.93 ± 0.08	-1.10 ± 0.11	-1.07 ± 0.05	-1.00 ± 0.11
	2	0.10 ± 0.02	0.09 ± 0.03	0.09 ± 0.05	0.16 ± 0.03	0.20 ± 0.04	0.14 ± 0.05	0.19 ± 0.06
	3	1.24 ± 0.13	1.32 ± 0.07	1.29 ± 0.13	1.28 ± 0.04	1.28 ± 0.05	1.24 ± 0.04	1.36 ± 0.07
	4	-2.06 ± 0.08	-2.09 ± 0.17	-2.06 ± 0.25	-1.41 ± 0.19	-0.94 ± 0.13	-0.97 ± 0.15	-1.34 ± 0.20

ABFE calculations for lysozyme. Four different simulation protocols were utilized: Protocol A, B, and C (see Table 1 for protocol details), as well as Protocol O, which was performed as a control group to evaluate the performance of our other protocol implementations. Protocol O consisted of 4.5 ns of production simulation time per λ-window. The gradient timeseries were then extracted, equilibrated with a 0.5 ns equilibration period, and decorrelated. For each protocol, the average ΔG_{bind} was computed as an average of the 39 independent calculations. The experimental ΔG_{bind} value of -5.52[5] kcal/mol was used to evaluate the protocol performance. The average ΔG_{bind} computed with Protocol A, B, C, and O were -5.31 ± 0.12 kcal/mol, -5.59 ± 0.14 kcal/mol, -5.46 ± 0.19 kcal/mol, and -5.36 ± 0.14 kcal/mol, respectively. MAEs and RMSEs of all four protocols are shown in Table 3. Note that values of batch sizes greater than one were calculated by resampling 1,000,000 samples of a given batch size with replacement and then averaging each sample.

Table 3. MAE, RMSE and Computational Savings of ABFE Simulations for lysozyme.

Simulation Protocol	Batch Size	MAE (kcal/mol)	RMSE (kcal/mol)
O	1	0.637	0.892
A	1	0.617	0.772
B	1	0.699	0.871
C	1	0.945	1.176
O	2	0.491	0.640
A	2	0.454	0.566
B	2	0.498	0.618
C	2	0.669	0.833
O	3	0.413	0.532
A	3	0.382	0.476
B	3	0.405	0.505
C	3	0.546	0.681
O	5	0.335	0.425
A	5	0.315	0.392
B	5	0.314	0.393
C	5	0.423	0.529
O	10	0.255	0.321
A	10	0.225	0.313
B	10	0.225	0.283
C	10	0.301	0.376

In general, the MAE and RMSE of the control group (Protocol O) was comparable to that of Protocol A and B across all batch sizes, whereas Protocol C was significantly less accurate. Interestingly, Protocol A outperformed Protocol O with all batch sizes in both MAE and RMSE. Protocol B managed the same feat with batch sizes of 3, 5 and 10 in MAE and all batch sizes in RMSE. Significant computational savings were achieved on all alchemical steps using either Protocols A, B, or C, with Protocol C achieving approximately 60% average savings or greater across all alchemical steps (see Fig. S8), albeit with a significant accuracy penalty. Protocol A was able to improve accuracy across the board when measured by RMSE while achieving average savings of approximately 30-45% depending on the alchemical step. Of the protocols tested, Protocol B offered the best compromise between cost and accuracy with a batch size of 1, with approximately a 50% reduction in computational cost with comparable accuracy to Protocol O. All protocols and batch sizes obtained average error within 1 kcal/mol, which is comparable to the corresponding error reported in other studies.[13,25,26]
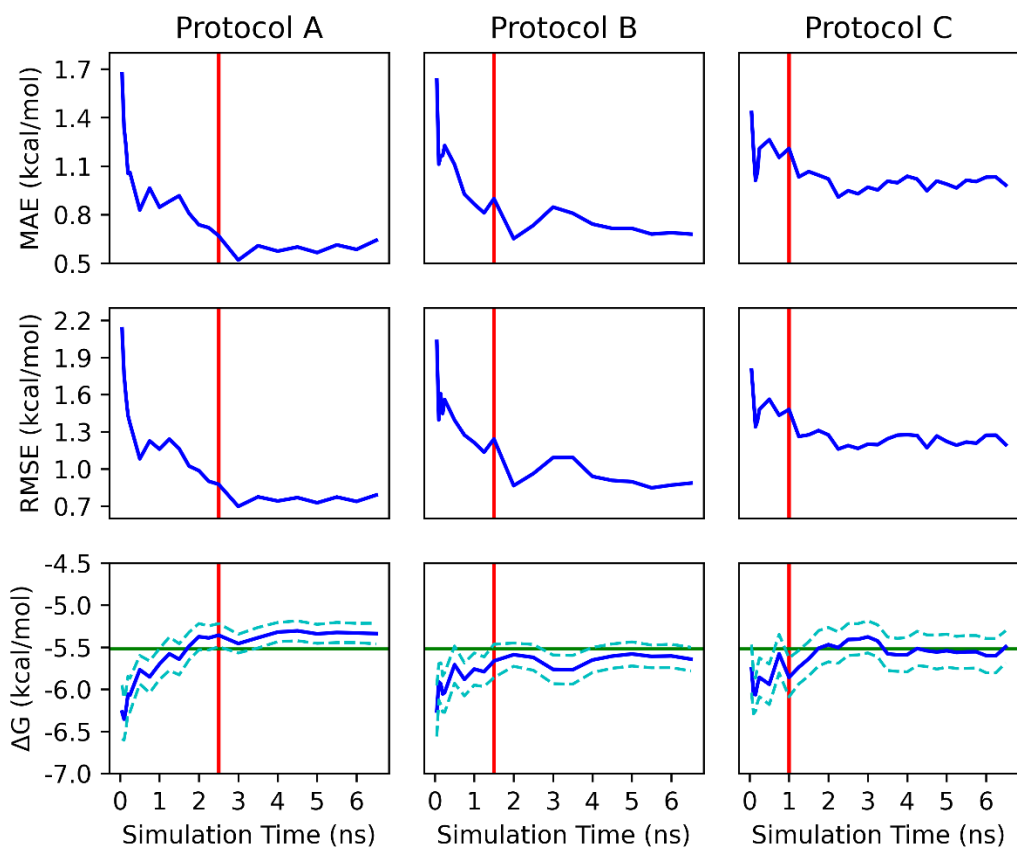
Figure 13. Lysozyme MAE, RMSE and overall ΔG_{bind} calculated from truncated gradients. The dashed line represents one standard error of the mean value, the red line represents the length of the initial simulation period, and the green line represents the experimental value of ΔG_{bind}.

As seen in Figure 13, the most significant accuracy gains were observed during the first nanosecond of production simulation time for Protocols A and B, whereas Protocol C displayed more muted gains. Protocol C displayed elevated inaccuracy at 1 ns of production simulation time when compared to Protocol A and B at the same time period, despite the fact that all protocols are equivalent at this point (see Table 1 for protocol details). This may indicate that Protocol C suffered from an uncommonly inaccurate batch of simulations, which may help explain its relative underperformance. All protocol average ΔG_{bind} converge within 1-2 ns to within a standard error of the experimental value, which indicates that shorter maximum simulation times may be employed while maintaining accuracy.

As opposed to the RBFE results, repeated runs had a significant impact on MAE and RMSE. Within Protocol O, MAE decreased from 0.637 kcal/mol to 0.230 kcal/mol when moving from a batch size of 1 to 10. Protocols A, B, and C showed similar trends, with MAE decreasing from 0.617 kcal/mol to 0.225 kcal/mol, from 0.699 kcal/mol to 0.225 kcal/mol, and from 0.945 to 0.301, respectively. The difference in MAE between Protocols A and B decreased from 0.082 kcal/mol to 0.000 kcal/mol as batch size increased from 1 to 10. Similar results were achieved when comparing RMSE values, with Protocol B displaying a three-fold RMSE decrease and the difference between Protocols A and B RMSE decreasing from 0.101 kcal/mol to 0.030 kcal/mol, with Protocol B becoming superior, as batch size increased from 1 to 10 replicates. These results suggest that in high-throughput virtual screening campaigns utilizing ensembles of ABFE simulations, Protocol A or B will outperform other protocols with uniform resource allocation. Furthermore, while Protocol A can achieve higher accuracy in one-off simulations, albeit with significantly higher cost, this advantage evaporates as batch size increases. At a batch size of 10 replicates, significant savings are realized with Protocol B with no relative accuracy penalty.

ABFE PLpro. Three 100 ns simulations were performed at each λ-window. The gradients of the short-run simulations were evaluated against those of the long-run simulations in an analogous manner as described for the RBFE simulations. Similarly to the RBFE PLpro study, the amount of simulation time applied to each λ-window varied by λ-window, simulation protocol, and

28

alchemical step (see Fig. S9). We note once again that λ-windows in the protein-ligand complex and restraint addition alchemical steps with elevated simulation times in Protocol A tended to have elevated simulation times in Protocols B and C, with the analogous pattern holding for Protocols B and C (see Table 1 for protocol details).

For all protocols, the protein-ligand complex and solvated ligand step short-run simulations, as well as the overall ΔG_{bind} short-run simulations, converged quickly towards their respective long-run averages and were within error after 1-2 ns of simulation time, as seen below in Figure 14. The restraint addition step short-run simulations, however, remained well outside of error. Analysis of this alchemical step showed that regardless of equilibration protocol employed, over 40 ns of production simulation time per λ-window is required to achieve the average values of approximately 2 kcal/mol (Fig. 15), which is significantly more resources than can be dedicated for this purpose.
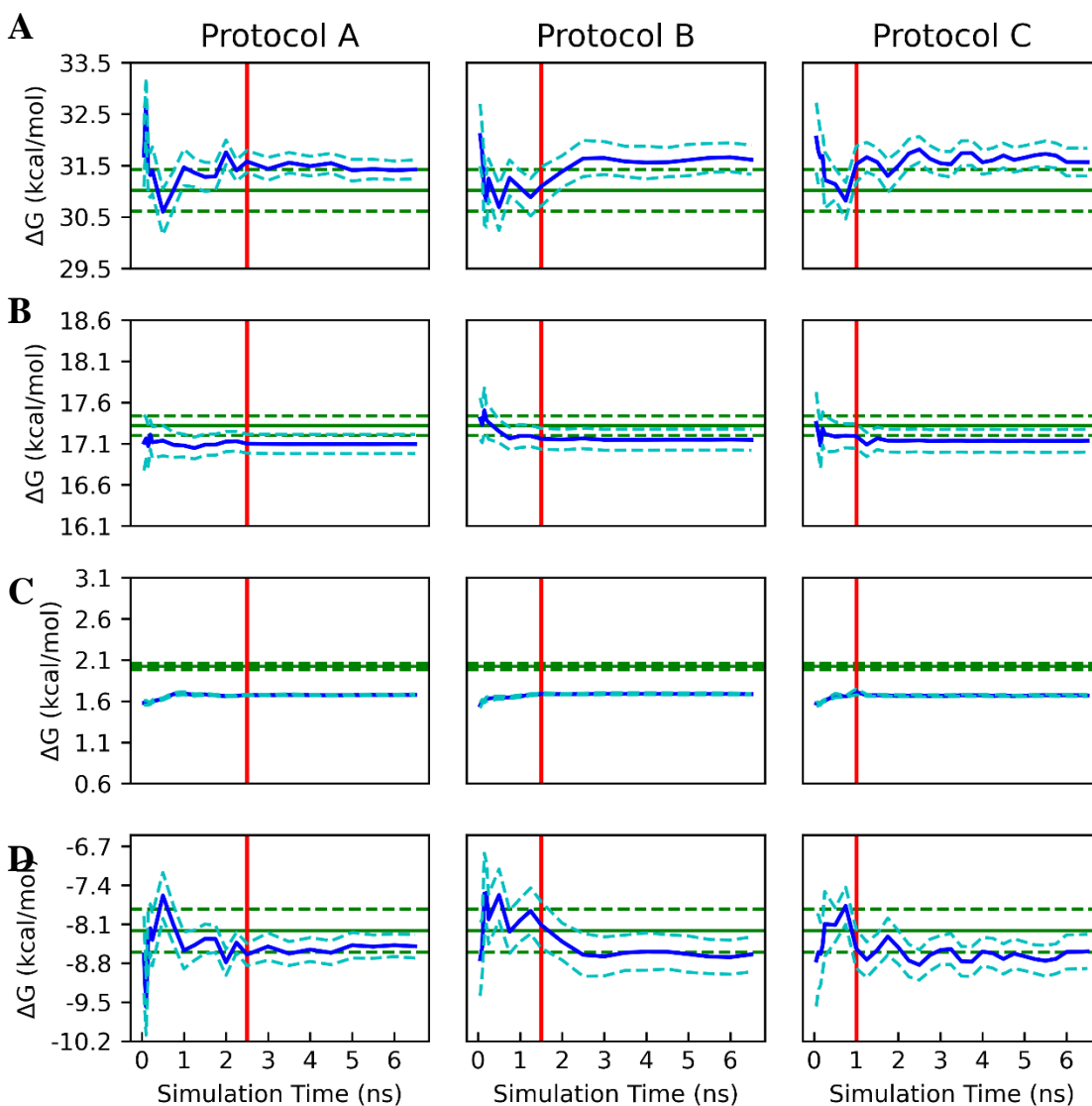
Figure 14. A) PLpro ABFE protein-ligand complex step ΔGs calculated from truncated gradients by protocol. B) PLpro ABFE solvated ligand step ΔGs calculated from truncated gradients by protocol. C) PLpro ABFE restraint addition step ΔGs calculated from truncated gradients by protocol. A) PLpro ABFE overall ΔG$_{bind}$ values calculated from truncated gradients by protocol. The blue solid line represents the mean short-run value, the blue dashed line represents one standard error of the mean short-run value, the green solid line represents the long-run mean value calculated with AED, the green dashed line represents one standard error of the long-run mean value, and the red line represents the length of the initial simulation period.
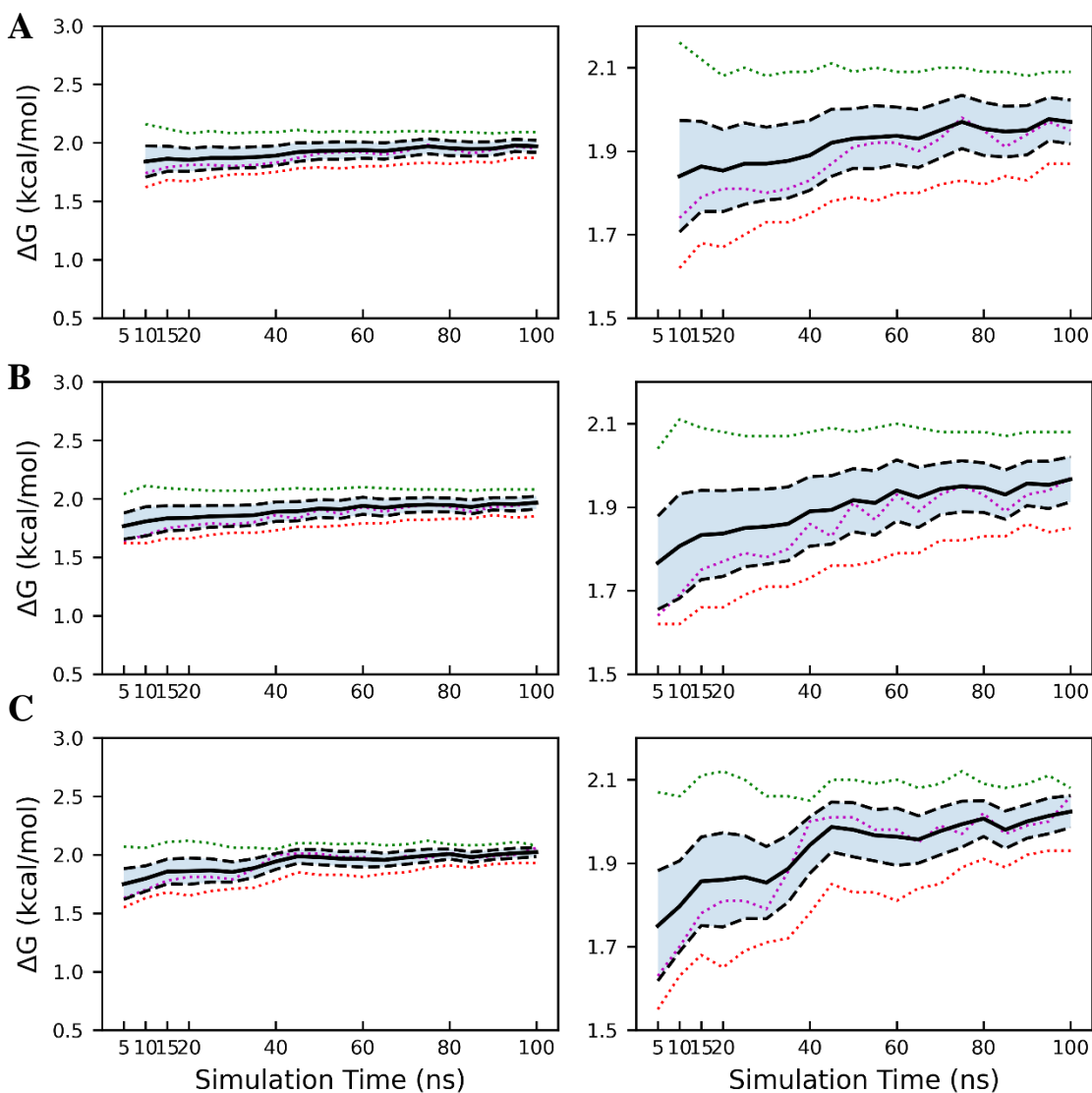
Figure 15. A) PLpro ABFE long-run restrain addition step ΔG calculated from truncated gradients with a 5 ns equilibration period. B) PLpro ABFE long-run restrain addition step ΔG calculated from truncated gradients with a 2 ns equilibration period. C) PLpro ABFE long-run restrain addition step ΔG calculated from truncated gradients with AED. The left column displays ΔG with a range of 2.5 kcal/mol, while the right column displays ΔG zoomed in. The dotted lines each represent an individual replicate while the black solid and dashed lines represent the average value and standard error, respectively. We note that the overall uncertainty was minimized with the AED equilibration protocol.

31

Average protein-ligand complex ΔG, solvated ligand ΔG, restraint addition ΔG and overall ΔG$_{bind}$ of the short-run simulations and long-run simulations with all three equilibration methods were computed and are tabulated in Table 4.

Table 4 ΔGs of Short-Run and Long-Run PLpro ABFE Simulations by Alchemical Step.

	Overall ΔG (kcal/mol)	Protein-Ligand Complex (kcal/mol)	Solvated Ligand (kcal/mol)	Restraint Addition (kcal/mol)	Protein-Ligand Complex Simulation Time (ns)	Solvated Ligand Simulation Time (ns)	Restraint Addition Simulation Time (ns)
Long-run AED Equilibration	-8.21 ± 0.47	31.02 ± 0.50	17.32 ± 0.15	2.02 ± 0.05	900	900	700
Long-run 2ns Equilibration	-7.65 ± 0.85	30.50 ± 0.91	17.33 ± 0.15	1.97 ± 0.07	900	900	700
Long-run 5ns Equilibration	-7.85 ± 0.67	30.71 ± 0.71	17.32 ± 0.16	1.97 ± 0.06	900	900	700
Protocol A	-8.50 ± 0.21	31.42 ± 0.18	17.10 ± 0.12	1.68 ± 0.01	27.8 ± 0.4	22.9 ± 0.1	26.5 ± 0.4
Protocol B	-8.64 ± 0.31	31.61 ± 0.28	17.15 ± 0.13	1.69 ± 0.01	20.1 ± 0.6	15.7 ± 0.2	19.2 ± 0.5
Protocol C	-8.59 ± 0.31	31.56 ± 0.27	17.14 ± 0.14	1.67 ± 0.01	16.7 ± 0.4	11.3 ± 0.2	14.1 ± 0.5

For both the protein-ligand complex and solvated ligand steps, there was not a significant difference between ΔGs obtained from the any of the short-run protocols and any of the long-run equilibration methods. While the restraint addition step did show significant deviation, the overall magnitude is small, and the errors are cancelled on average in the other alchemical steps. However, computational cost did vary across short-run protocols, with Protocol C converging in approximately 50-60% of the simulation time allocated to Protocol A depending on the alchemical step, with Protocol B falling between the two but closer to C.

Conclusion

We have presented a data driven procedure for the optimization of computational resources usage for both ABFE and RBFE calculations with thermodynamic integration. Our RBFE scheme affords up to 85% computational resource reduction when compared to the CDK2 benchmark system results published by Song *et al*, while maintaining average MAE of approximately 1 kcal/mol. Our protocols have successfully approximated long-run simulations of small RBFE mutations performed on the PLpro system, with the larger ligand 4 mutation deviating more significantly but still within 1 kcal/mol on average. Our ABFE schemes yield fast one-off calculations with similar accuracy when compared to a base case of uniform and constant resource allocation on the T4

Lysozyme L99A/M102Q mutant in complex with N-phenylglycinonitrile, and several implementations become more accurate while maintaining computational efficiency as batch size increases. ABFE PLpro simulations displayed strong agreement between long-run 100 ns simulations and short-run simulations, with no significant deviation observed in the protein-ligand complex step, solvated ligand step, or overall computed ΔG_{bind}.

For future high-throughput RBFE campaigns, we recommend dividing mutations into two groups: "easy" mutations, consisting of those with few changes in heavy atoms or rotatable bonds, and "difficult" mutations, consisting of those with many changes in heavy atoms and rotatable bonds. For the easy mutations, we recommend performing one-off simulations using very short protocols (Protocols C or D) as these have been shown to be just as accurate on both the CDK2 and PLpro systems while achieving significant computational savings. For more difficult mutations, longer protocols are appropriate (Protocols A or E), as these require more sampling time to account for the larger amount of phase space available to the ligand. For future high-throughput ABFE campaigns, we recommend using 3-5 replicates with Protocol B. If this is too expensive, Protocols C or D, or even shorter protocols can be employed; however, prior testing should be employed to ensure they are maintaining the requisite accuracy. In general, savings can be achieved by limiting the production simulation time of the solvated ligand and restraint addition steps, whereas resources should be concentrated on the protein-ligand complex step. This could be achieved by employing a mixed protocol utilizing Protocols A or B for the protein-ligand complex and Protocols C or D for the solvated ligand and restraint addition.

Supporting Information

- Schematic representation of virtual bond approach for protein-ligand restraints.
- Average MAE, RMSE, and computational savings by alchemical step of RBFE simulations for the CDK2 benchmark system.
- Representative RMSD plots for a long-run PLpro simulation.
- Plots of average $\Delta\Delta G_{bind}$ and its components for PLpro ligand 2 and 3 calculated from truncated trajectories.
- Plots of $dV/d\lambda$ convergency and dihedral angles distribution for PLpro ligand 4.
- Average total simulation time of ABFE simulations for Lysozyme and PLpro systems.

Author Information

Corresponding Author

Maria G Kurnikova – Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, United States; orcid.org/0000-0002-8010-8374;
Email: kurnikova@cmu.edu

Authors

S. Benjamin Koby – Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, United States; orcid.org/0009-0007-6643-7271

Evgeny Gutkin – Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, United States; orcid.org/0000-0003-4522-6049

Shree Patel – Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, United States

Author contributions

Notes

The authors declare no competing financial interest.

Acknowledgements

References

(1)   *Biomolecular Simulations*; Monticelli, L., Salonen, E., Eds.; Humana Press: Totowa, NJ, 2013; Vol. 924. https://doi.org/10.1007/978-1-62703-017-5.

(2)   Gallicchio, E.; Levy, R. M. Recent Theoretical and Computational Advances for Modeling Protein–Ligand Binding Affinities. In *Advances in Protein Chemistry and Structural Biology*; Christov, C., Ed.; Academic Press, 2011; Vol. 85, pp 27–80. https://doi.org/https://doi.org/10.1016/B978-0-12-386485-7.00002-8.

(3)   Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, 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* (11), 5595–5623. https://doi.org/10.1021/acs.jcim.0c00613.

(4)   McDonnell, J. M. Surface Plasmon Resonance: Towards an Understanding of the Mechanisms of Biological Molecular Recognition. *Curr Opin Chem Biol* **2001**, *5* (5), 572–577. https://doi.org/10.1016/S1367-5931(00)00251-9.

(5)   Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu Rev Biophys* **2017**, *46* (1), 531–558. https://doi.org/10.1146/annurev-biophys-070816-033654.

(6)   Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W. Rigorous Free Energy Simulations in Virtual Screening. *J Chem Inf Model* **2020**, *60* (9), 4153–4169. https://doi.org/10.1021/acs.jcim.0c00116.

(7)   Gusev, F.; Gutkin, E.; Kurnikova, M. G.; Isayev, O. Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling. *J Chem Inf Model* **2023**, *63* (2), 583–594. https://doi.org/10.1021/acs.jcim.2c01052.

(8)   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* (12), 2911–2937. https://doi.org/10.1021/acs.jcim.7b00564.

(9)   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* (7), 2695–2703. https://doi.org/10.1021/ja512751q.

(10)   Song, L. F.; Merz, K. M. Evolution of Alchemical Free Energy Methods in Drug Discovery. *J Chem Inf Model* **2020**, *60* (11), 5308–5318. https://doi.org/10.1021/acs.jcim.0c00547.

(11)   Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.

(12)   Guterres, H.; Im, W. Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations. *J Chem Inf Model* **2020**, *60* (4), 2189–2198. https://doi.org/10.1021/acs.jcim.0c00057.

(13)   Chen, W.; Cui, D.; Jerome, S. V.; Michino, M.; Lenselink, E. B.; Huggins, D. J.; Beautrait, A.; Vendome, J.; Abel, R.; Friesner, R. A.; Wang, L. Enhancing Hit Discovery in Virtual Screening through Absolute Protein–Ligand Binding Free-Energy Calculations. *J Chem Inf Model* **2023**, *63* (10), 3171–3185. https://doi.org/10.1021/acs.jcim.3c00013.

(14)   Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin Drug Discov* **2015**, *10* (5), 449–461. https://doi.org/10.1517/17460441.2015.1032936.

(15)   Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem Rev* **2019**, *119* (16), 9478–9508. https://doi.org/10.1021/acs.chemrev.9b00055.

(16)   Bhati, A. P.; Wan, S.; Coveney, P. v. Ensemble-Based Replica Exchange Alchemical Free Energy Methods: The Effect of Protein Mutations on Inhibitor Binding. *J Chem Theory Comput* **2019**, *15* (2), 1265–1277. https://doi.org/10.1021/acs.jctc.8b01118.

(17)   He, X.; Liu, S.; Lee, T.-S.; Ji, B.; Man, V. H.; York, D. M.; Wang, J. Fast, Accurate, and Reliable Protocols for Routine Calculations of Protein–Ligand Binding Affinities in Drug Design Projects Using AMBER GPU-TI with Ff14SB/GAFF. *ACS Omega* **2020**, *5* (9), 4611–4619. https://doi.org/10.1021/acsomega.9b04233.

(18)   Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J Comput Phys* **1976**, *22* (2), 245–268. https://doi.org/10.1016/0021-9991(76)90078-4.

(19)   Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J Chem Phys* **2008**, *129* (12), 124105. https://doi.org/10.1063/1.2978177.

(20)   Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J Chem Phys* **1935**, *3* (5), 300–313. https://doi.org/10.1063/1.1749657.

(21)   Paliwal, H.; Shirts, M. R. A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods. *J Chem Theory Comput* **2011**, *7* (12), 4115–4134. https://doi.org/10.1021/ct2003995.

(22) Wade, A. D.; Bhati, A. P.; Wan, S.; Coveney, P. v. Alchemical Free Energy Estimators and Molecular Dynamics Engines: Accuracy, Precision, and Reproducibility. *J Chem Theory Comput* **2022**, *18* (6), 3972–3987. https://doi.org/10.1021/acs.jctc.2c00114.

(23) Song, L. F.; Lee, T.-S.; Zhu, C.; York, D. M.; Merz, K. M. Using AMBER18 for Relative Free Energy Calculations. *J Chem Inf Model* **2019**, *59* (7), 3128–3135. https://doi.org/10.1021/acs.jcim.9b00105.

(24) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, 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* (11), 5595–5623. https://doi.org/10.1021/acs.jcim.0c00613.

(25) Khalak, Y.; Tresadern, G.; Aldeghi, M.; Baumann, H. M.; Mobley, D. L.; de Groot, B. L.; Gapsys, V. Alchemical Absolute Protein–Ligand Binding Free Energies for Drug Design. *Chem Sci* **2021**, *12* (41), 13958–13971. https://doi.org/10.1039/D1SC03472C.

(26) Fu, H.; Chen, H.; Cai, W.; Shao, X.; Chipot, C. BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *J Chem Inf Model* **2021**, *61* (5), 2116–2123. https://doi.org/10.1021/acs.jcim.1c00269.

(27) Fu, H.; Zhou, Y.; Jing, X.; Shao, X.; Cai, W. Meta-Analysis Reveals That Absolute Binding Free-Energy Calculations Approach Chemical Accuracy. *J Med Chem* **2022**, *65* (19), 12970–12978. https://doi.org/10.1021/acs.jmedchem.2c00796.

(28) Chen, H.; Maia, J. D. C.; Radak, B. K.; Hardy, D. J.; Cai, W.; Chipot, C.; Tajkhorshid, E. Boosting Free-Energy Perturbation Calculations with GPU-Accelerated NAMD. *J Chem Inf Model* **2020**, *60* (11), 5301–5307. https://doi.org/10.1021/acs.jcim.0c00745.

(29) Harger, M.; Li, D.; Wang, Z.; Dalby, K.; Lagardère, L.; Piquemal, J.; Ponder, J.; Ren, P. Tinker-OpenMM: Absolute and Relative Alchemical Free Energies Using AMOEBA on GPUs. *J Comput Chem* **2017**, *38* (23), 2047–2055. https://doi.org/10.1002/jcc.24853.

(30) Kutzner, C.; Kniep, C.; Cherian, A.; Nordstrom, L.; Grubmüller, H.; de Groot, B. L.; Gapsys, V. GROMACS in the Cloud: A Global Supercomputer to Speed Up Alchemical Drug Design. *J Chem Inf Model* **2022**, *62* (7), 1691–1711. https://doi.org/10.1021/acs.jcim.2c00044.

(31) Abel, R.; Wang, L.; Mobley, D. L.; Friesner, R. A. A Critical Review of Validation, Blind Testing, and Real- World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Curr Top Med Chem* **2017**, *17* (23), 2577–2585. https://doi.org/10.2174/1568026617666170414142131.

(32) Cruz, J.; Wickstrom, L.; Yang, D.; Gallicchio, E.; Deng, N. Combining Alchemical Transformation with a Physical Pathway to Accelerate Absolute Binding Free Energy Calculations of Charged Ligands to Enclosed Binding Sites. *J Chem Theory Comput* **2020**, *16* (4), 2803–2813. https://doi.org/10.1021/acs.jctc.9b01119.

(33) Deng, N.; Cui, D.; Zhang, B. W.; Xia, J.; Cruz, J.; Levy, R. Comparing Alchemical and Physical Pathway Methods for Computing the Absolute Binding Free Energy of Charged Ligands. *Physical Chemistry Chemical Physics* **2018**, *20* (25), 17081–17092. https://doi.org/10.1039/C8CP01524D.

(34) Muegge, I.; Hu, Y. Recent Advances in Alchemical Binding Free Energy Calculations for Drug Discovery. *ACS Med Chem Lett* **2023**. https://doi.org/10.1021/acsmedchemlett.2c00541.

(35) 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* (6), 3120–3130. https://doi.org/10.1021/acs.jcim.0c00165.

(36) Khalak, Y.; Tresadern, G.; Hahn, D. F.; de Groot, B. L.; Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J Chem Theory Comput* **2022**, *18* (10), 6259–6270. https://doi.org/10.1021/acs.jctc.2c00752.

(37) Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S. Reaction-Based Enumeration, Active Learning, and Free Energy Calculations To Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin-Dependent Kinase 2 Inhibitors. *J Chem Inf Model* **2019**, *59* (9), 3782–3793. https://doi.org/10.1021/acs.jcim.9b00367.

(38) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating High-Throughput Virtual Screening through Molecular Pool-Based Active Learning. *Chem Sci* **2021**, *12* (22), 7866–7881. https://doi.org/10.1039/D0SC06805E.

(39) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent Sci* **2020**, *6* (6), 939–949. https://doi.org/10.1021/acscentsci.0c00229.

(40) Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B. K.; Jerome, S. v. Efficient Exploration of Chemical Space with Docking and Deep Learning. *J Chem Theory Comput* **2021**, *17* (11), 7106–7119. https://doi.org/10.1021/acs.jctc.1c00810.

(41) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Sci Data* **2020**, *7* (1), 134. https://doi.org/10.1038/s41597-020-0473-z.

(42) Kanada, R.; Tokuhisa, A.; Tsuda, K.; Okuno, Y.; Terayama, K. Exploring Successful Parameter Region for Coarse-Grained Simulation of Biomolecules by Bayesian Optimization and Active Learning. *Biomolecules* **2020**, *10* (3), 482. https://doi.org/10.3390/biom10030482.

(43) Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing Active Learning for Free Energy

Calculations. *Artificial Intelligence in the Life Sciences* **2022**, *2*, 100050. https://doi.org/10.1016/j.ailsci.2022.100050.

(44) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining Docking and Molecular Dynamic Simulations in Drug Design. *Med Res Rev* **2006**, *26* (5), 531–568. https://doi.org/10.1002/med.20067.

(45) Graves, A. P.; Shivakumar, D. M.; Boyce, S. E.; Jacobson, M. P.; Case, D. A.; Shoichet, B. K. Rescoring Docking Hit Lists for Model Cavity Sites: Predictions and Experimental Testing. *J Mol Biol* **2008**, *377* (3), 914–934. https://doi.org/10.1016/j.jmb.2008.01.049.

(46) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J Phys Chem B* **2003**, *107* (35), 9535–9551. https://doi.org/10.1021/jp0217839.

(47) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput Phys Commun* **1995**, *91* (1–3), 1–41. https://doi.org/10.1016/0010-4655(95)00041-D.

(48) le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without Compromise—A Mixed Precision Model for GPU Accelerated Molecular Dynamics Simulations. *Comput Phys Commun* **2013**, *184* (2), 374–380. https://doi.org/10.1016/j.cpc.2012.09.022.

(49) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* **2012**, *8* (5), 1542–1555. https://doi.org/10.1021/ct200909j.

(50) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* **2013**, *9* (9), 3878–3888. https://doi.org/10.1021/ct400314y.

(51) Beckstein, O.; Dotson, D.; Wu, Z.; Wille, D.; Marson, D.; Kenney, I.; Shuail; Lee, H.; trje3733; Lim, V.; Schlaich, A.; Alibay, I.; Henin, J.; Barhaghi, M. S.; Merz, P.; Joseph, T.; Hsu, W.-T. Alchemistry/Alchemlyb: 2.0.0 (2.0.0). 2022.

(52) Shirts, M.; Beauchamp, K.; Naden, L.; Chodera, J.; Rodriguez-Guerra, J.; Martiniani, S.; Stern, C.; Henry, M.; Fass, J.; Gowers, R.; McGibbon, R. T.; Dice, B.; Jones, C.; Dotson, D.; Burgin, T. Choderalab/Pymbar: 3.1.1 (3.1.1). 2022.

(53) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/ct400341p.

(54) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J Comput Chem* **2005**, *26* (16), 1668–1688. https://doi.org/10.1002/jcc.20290.

(55) GAFF and GAFF2 are public domain force fields and are part of the AmberTools16 distribution, available for download at http://amber.org (accessed October 2018).According to the AMBER development team, the improved version of GAFF, GAFF2, is an ongoing project aimed at "reproducing both the high quality interaction energies and key liquid properties such as density, heat of vaporization and hydration free energy". GAFF2 is expected "to be an even more successful general purpose force field and that GAFF2-based scoring functions will significantly improve the successful rate of virtual screenings".

(56) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model*; 1993; Vol. 97. https://pubs.acs.org/sharingguidelines.

(57) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr. , J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision A.02. Gaussian, Inc: Wallingford, CT 2009.

(58) 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.; Kittao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, Jr. , J. A.; 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 2019.

(59) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **1979**, *7* (1), 1–26.

(60) Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J Chem Theory Comput* **2016**, *12* (4), 1799–1805. https://doi.org/10.1021/acs.jctc.5b00784.

(61) Lamberti, P. W.; Majtey, A. P.; Madrid, M.; Pereyra, M. E. Jensen-Shannon Divergence: A Multipurpose Distance for Statistical and Quantum Mechanics. In *AIP Conference Proceedings*; AIP, 2007; pp 32–37. https://doi.org/10.1063/1.2746720.