Alchemical Free Energy Estimators and Molecular Dynamics Engines: Accuracy, Precision and Reproducibility

Alexander D. Wade,[†] Agastya P. Bhati,[†] Shunzhou Wan,[†] and Peter V. Coveney*,[†,‡]

[†] *Centre for Computational Science, Department of Chemistry, University College London, London UK*

[‡] *Informatics Institute, University of Amsterdam, Amsterdam, Netherlands.*

E-mail: p.v.coveney@ucl.ac.uk

Phone: +44 (0)20 7679 4560

ORCIDs: Alexander D. Wade - 0000-0003-1500-3733 , Agastya P. Bhati - 0000-0003-4539-4819, Peter V. Coveney - 0000-0002-8787-7256

Abstract

The binding free energy between a ligand and its target protein is an essential quantity to know at all stages of the drug discovery pipeline. Assessing this value computationally can offer insight into where efforts should be focused in the pursuit of effective therapeutics to treat myriad diseases. In this work we examine the computation of alchemical relative binding free energies with an eye to assessing reproducibility across popular molecular dynamics packages and free energy estimators. The focus of this work is on 54 ligand transformations from a diverse set of protein targets: MCL1, PTP1B, TYK2, CDK2 and thrombin. These targets are studied with three popular molecular dynamics packages: OpenMM, NAMD2 and NAMD3. Trajectories collected with these packages are used to compare relative binding free energies calculated with thermodynamic integration and free energy perturbation methods. The resulting binding free energies show good agreement between molecular dynamics packages with an average mean unsigned error between packages of 0.5 $kcal/mol$ The correlation between packages is very good with the lowest Spearman's, Pearson's and Kendall's tau correlation coefficient between two packages being 0.91, 0.89 and 0.74 respectively. Agreement between thermodynamic integration and free energy perturbation is shown to be very good when using ensemble averaging.

1 Introduction

When applied rigorously, computational free energy methods offer the ability to make accurate and precise predictions for protein-ligand binding affinities[1]. Physics based free energy methods, whilst historically being prohibitively expensive, have now become routine, with the development of GPU hardware and GPU accelerated molecular dynamics (MD) codes[2–4]. The way in which these calculations are structured provides many opportunities for concurrent execution across high performance computers, allowing predictions for binding affinities including reliable error estimates to be made in the order of hours, a critical time frame in the domains of drug design and personalized medicine.

The accuracy of these calculations has been improved over time as the force fields, used to parameterize the system, have seen continued development[5–8]. Ongoing work in the field aims to further these improvements with large collaborative endeavors such as the Open Force Field Initiative[9] and the development of systematic methods for force field optimization such as Force Balance[10]. Being able to calculate these binding free energies accurately can be of significant benefit to drug design campaigns, helping to reduce the large cost involved with drug development[11]. Moreover, these calculations can allow for much larger areas of chemical space to be explored than would be possible experimentally. Compounds drawn from this chemical space can be selected from numerous sources such as chemical libraries[12], repurposing of approved drugs[13], using generative AI methods[14–16] or even other free energy calculations[17].

Another aspect of MD-based free energy calculations is the extreme sensitivity of such calculations to their initial conditions[18]. It has been shown that free energies derived from two independent MD simulations, only varying in their starting velocities, can vary by a substantial amount; the exact figure depends on the type of method used and the system studied[19–26]. MD-based free energy methods require ensemble averaging across the conformations generated. However, the practice has been to perform time averaging over a single trajectory relying on the ergodic theorem which equates time averaging to ensemble averaging. It is worth mentioning that the ergodic theorem holds true only in the limit of infinite time which is far from the typical length of simulations performed. This explains the observed differences in free energies between repeat simulations. Indeed, a recent study showed that ensembles are required to handle both aleatoric and parametric uncertainty in MD simulation[27]. It has also been shown that running an ensemble of independent simulations varying only in their starting conditions, or in other words, an ensemble simulation yields precise and reproducible results[21,25]. In particular, methods such as ESMACS[21,22] and TIES[25,26] have been developed based on such ensemble simulations so as to ensure reproducibility and hence reliability of the predicted free energies. Recently, we have also shown that the distributions of free energies obtained from such ensemble simulations are in general not Gaussian, as is the common assumption; rather they exhibit non-normality which has interesting and important consequences[28,29].

Alchemical binding free energy calculations are a class of free energy method which involve the transformation of one or more chemical moieties in the system to another[30]. Alchemical protein-ligand binding calculations can be performed in an absolute or relative fashion[31,32]. In an absolute calculation the binding free energy of a ligand is calculated by completely removing the ligand from the protein-ligand complex. Alternatively one can perform relative calculations which compare the binding free energy between two ligands. During a relative calculation one ligand is transformed, via unphysical "alchemical" intermediate states, into another. The two ligands studied generally have a highly conserved chemical structure; this is both a strength and a weakness of the method, since the practitioner is restricted to study cogeneric ligands but benefits from potential gains in accuracy and precision resulting from studying smaller alchemical changes when compared to absolute methods. Cogeneric ligands also come with some other tangential benefits such as avoiding complicating factors such as standard states corrections[33,34]. In this study, we have employed the ensemble simulation based TIES[25] to correctly handle the uncertainties associated with such calculations and extended this to apply to free energy perturbation methods.

In this work we consider only relative binding free energy (RBFE) calculations. Several existing software applications can facilitate these calculations such as PMX based on GROMACS[35], FEP+ proprietary software produced by Schrödinger[36] or FESetup[37]. Our group has recently publicly released the comprehensive TIES toolkit[38] to automatically setup, execute and analyse such calculations; this software was used to set up and perform all calculations for this study. The specifics of RBFE methodologies vary between implementations, being based on user choices about how to carry out calculations. Some key areas where this variation could significantly influence the results include the topology of the transformed moieties, the thermodynamic path followed between end states and how much sampling is performed in each state. These factors introduce some uncertainty in the results but this is generally controlled by probing them on a case by case basis.

For the application of RBFE calculations to the protein-ligand binding problem, one aspect of uncertainty quantification which has received less attention is the variation in results across MD packages. In a previous work by Rizzi *et al.* a wide array of alchemical methods were compared, including potential of mean force and weighted ensemble methods[39]. This study reported that the variability in the absolute binding free energy across the methods tested is in the range of 0.3 to 1.0 $kcal/mol$. However, due to differences between the methods tested comparisons are difficult to draw across alchemical methods or estimators. Moreover, unlike our current study, this study does not directly compare the performance of different MD packages using the same alchemical methods, which adds too many variables for systematic and direct comparisons to be made. The input systems used by Rizzi *et al.* were closely matched but with some differences arising from factors such as different Coulomb constants used by AMBER and CHARMM or differing implementations of particle-mesh Ewald methods or Lennard-Jones (LJ) cutoff schemes. Technical differences between MD codes is a recurring issue which complicates the comparison of calculations and plays an important role in the present study.

Another study which performs a comparison between estimators using some simple benchmark systems has been carried out previously by Paliwal *et al.*[40] who study in detail the properties of numerous perturbative estimators as well as thermodynamic integration. All estimators are examined using GROMACS, allowing for meaningful comparisons to be made. However, the systems used by Paliwal *et al.* are small toy models and hence are not relevant for larger protein-ligand systems as used in this work. One of the ways in which uncertainty is quantified in their work was to run an ensemble of 100 simulations and calculate the mean and standard deviation of the binding free energies from each simulation. Using large ensembles allowed Paliwal *et al.* to quantify the type of distribution for calculated hydration free energies. From the calculated binding free energy distributions, it is concluded that the assumptions of Gaussian distributed errors in free energies are usually valid for most methods studied. This is contrary to the observations made in our work where, when using the same free energy estimators for an investigation of more complex protein-ligand systems, it is found that Gaussian distributions cannot be assumed, as also reported in some of our previous studies[28,29].

In the present paper we investigate the reproducibility of relative binding free energy calculations using three MD packages and two free energy estimators. Use of ensemble based simulations will be made to control uncertainties, as is essential for any calculation reliant on chaotic MD trajectories. Using ensembles to provide robust error control, we aim to identify statistically significant differences in the results from the different MD packages and estimators.

2 Theory

In this section, we outline the essential theory underpinning the alchemical methods we study.

2.1 Alchemical Methods

Applied to protein-ligand binding problems, alchemical methods involve changing chemical moieties in the studied system and calculating the free energy differences associated with these changes. Since in atomistic simulations systems are parameterized by force-fields, the transformation of chemical moieties can be achieved by modifying the atomistic parameters of the system. The variable λ is designated to control the modified parameters of the system, turning on and off relevant inter and intra molecular potentials. The reduced potential $u(\boldsymbol{x}, \lambda)$ of such a system can therefore be written as a function of the controlling parameter,

$$u(\boldsymbol{x}, \lambda) = \frac{1}{k_B T} \left[U(\boldsymbol{x}, \lambda) + pV(\boldsymbol{x}) \cdots \right]. \tag{1}$$

Here $\boldsymbol{x}$ is the configuration of the system, U is the potential energy, p the pressure and V the volume, plus any other terms relevant to the specific ensemble in which the simulation is run eg. NPT, grand, etc. As an example, consider the transformation of some ligand A to some ligand B. The value of λ ranges between 0 and 1, with a λ of 0 the system is in a state describing ligand A and with a λ of 1 the system is in a state describing ligand B. Typically λ will take multiple intermediate values between the end states 0 and 1; this range of λ values $\lambda_0, \lambda_{0.1}, \lambda_{0.2}...\lambda_1$ defines a set of alchemical states and simulations are performed in all these states. The choice of these states is not arbitrary and can affect both the accuracy and precision of the results.

2.2 Thermodynamic Integration

To calculate free energy differences with alchemical methods one of many available estimators can be used. In this work an application of thermodynamic integration (TI) estimator is made with enhanced sampling (TIES[25]); this methodology has been used in numerous studies to calculate accurate and precise RBFEs[25,26,29]. Centrally, TIES is based on the formally exact TI equation,

$$\Delta G = \int_0^1 \left\langle \frac{\partial u(\lambda, x)}{\partial \lambda} \right\rangle_\lambda \partial \lambda. \tag{2}$$

Here G is the Gibbs free energy and ΔG is the change in Gibbs free energy between two states A and B. ΔG is calculated by integrating $\left\langle \frac{\partial u(\lambda,x)}{\partial \lambda} \right\rangle_\lambda$ over the range of λ and this integration is performed numerically. It is worth highlighting that equation 2 is only strictly valid in the thermodynamic limit, when both left-hand-side and right-hand-side terms are unique numbers with no fluctuations. However, practically speaking, we work with finite systems and sample only a fraction of the full conformational space, which makes these quantities stochastic variables[18,28]. This implies that both the free energy as well as its derivative will have a distribution of values. Therefore, it is necessary to get the expectation value of these quantities using ensemble methods. The brackets $\langle . \rangle_\lambda$ denote an ensemble average in a thermodynamic state defined by the value of λ. To compute this ensemble average the configurations of particles can be sampled using Monte Carlo or MD methods and from these sampled configurations values of the potential are calculated and averaged. The traditional approach has been to perform a single "long" MD simulation to proxy ensemble averaging with time averaging. However, as discussed already, this is not reliable due to the extreme sensitivity of results obtained, arising from the initial conditions which are controlled by the random seeds used to initiate simulations. Thus, ensemble simulations are required to generate the ensemble of conformations in order to estimate an average and distribution of the calculated ΔG. In this work, the same idea of performing ensemble simulation to get the expectation value of the distribution of ΔG by performing stochastic integration of the distributions of $\left\langle \frac{\partial u(\lambda,x)}{\partial \lambda} \right\rangle_\lambda$ is applied using the TIES methodology.

2.3 Free Energy Perturbation

Parallel to the set based on TI are perturbative methods such as free energy perturbation (FEP) methods. The simplest estimators belonging to this class of perturbative methods are those based on the Zwanzig relation. However, it is known in FEP calculations that free energy estimates from the Zwanzig relation can be prone to bias stemming from the dominant contribution of rare samples when using finite sampling[41]. As such there exists several methods which aim to improve on the exponential averaging estimator; these are the Bennet acceptance ratio (BAR) and the Multistate Bennet acceptance ratio (MBAR). In this work the MBAR estimator is used, the derivation of this method is given in detail in the work of Shirts and Chodera *et al.*[42]. Here we present the relevant equation from this previous work for computing dimensionless free energies in a system with K total λ states,

$$f_i = -\ln \sum_{n=1}^{N} \frac{\exp(-u_i(\boldsymbol{x}_n))}{\sum_{k=1}^{K} N_k \exp(f_k - u_k(\boldsymbol{x}_n))}. \tag{3}$$

In this equation, $f_{i/k}$, is the dimensionless free energy for the state with $\lambda = \lambda_{i/k}$, N is total number of samples indexed by n, with N_k is the number of sample collected in state $\lambda = \lambda_k, u_{i/k}(\boldsymbol{x}_n)$ is then the reduced potential energy evaluated in state $\lambda = \lambda_{i/k}$ calculated using the configuration sampled in iteration n. Note that the summations run over all alchemical windows and thus information from all windows is combined to produce a free energy estimate; if only two windows are considered MBAR reduces to BAR[42]. This equation can be solved self-consistently with many solvers and these methods are implemented in the *pymbar* package[42], which was used in this work to compute results with MBAR.

The dimensionless free energies in equation 3 are combined into free energy differences and converted to the Gibbs free energies as follows,

$$\Delta f(\lambda_i, \lambda_i + 1) = f_{i+1} - f_i, \tag{4a}$$

$$\Delta G = k_B T \sum_{i=1}^{K-1} \Delta f(\lambda_i, \lambda_{i+1}). \tag{4b}$$

If the overlap in phase space between adjacent alchemical states is low it can be difficult to sample sufficiently to calculate trustworthy free energy differences with FEP methods[43]. No rigorous criteria exist which relates the expected variance in the calculated free energy to the amount of sampling or overlap between states for a given system. As a result there are numerous other ways in which the reliability of FEP calculations are tested[41,43], such as calculating convergence of results with the amount of sampling/number of alchemical windows or computing overlap distributions and overlap matrices[43]. The main way in which the variance in FEP calculation will be addressed in this work is through the use of ensembles of simulations. As described above for the TI estimator, the concept of ensemble simulations to obtain the expectation value of ΔG along with associated uncertainty will also be applied to FEP.

2.4 Ligand Protein Binding Free Energy

ΔG can be calculated with many different estimators. In order to calculate the binding free energy of ligand to protein, $\Delta\Delta G$, calculated values for ΔG are combined through a thermodynamic cycle[31]. In the case of RBFE for protein ligand binding the following thermodynamic cycle is routinely used,

$$\Delta\Delta G = \Delta G^{binding}_{L_B} - \Delta G^{binding}_{L_A} = \Delta G^{alch}_{complex} - \Delta G^{alch}_{solvent}; \tag{5}$$

here $\Delta G^{alch}_{solvent/complex}$ are the ΔGs calculated in equations 4 and 2 in the solvent/complex simulations (transforming L_A into L_B). Where the solvent simulation is the ligand in solvent and the complex simulation is the ligand in complex with the solvated protein. $\Delta G^{binding}_{L_A/L_B}$ is the binding free energy of ligand A/B to the protein. The difference of these alchemical free energies is equal to the difference of binding free energy of the ligands A and B which allows for the final $\Delta\Delta G$ to be calculated.

3 Methods

The RBFE calculations performed in this work are calculated using three molecular dynamics packages; these are NAMD2, NAMD3 and OpenMM. OpenMM can perform MD calculations on multiple platforms (CPU, CUDA, OpenCL); in this work all OpenMM calculations are performed using the CUDA platform with OpenMM 7.4.2. The NAMD2 calculations are performed on CPUs and NAMD 3.08 alpha calculations are performed on GPUs.

To automate the setup and running of these simulations we have developed and released an open source Python package called TIES MD which is available online [1]. This study uses the existing input files from previous research which works with TIES MD; novel input ligand transformations can be generated using an online service, TIES 20 [2]. The combination of TIES MD and TIES 20 allows anyone to freely and easily use the TIES protocol to calculate binding free energies.

Our simulations were run across several high performance computers including Summit at the Oak Ridge National Laboratory (ORNL) and SuperMUC-NG at the Leibniz Super Computing Centre. The performance of OpenMM and NAMD3 can be compared simply using the same hardware and for example running on one Nvidia V100 GPU with a ligand-protein complex of 35k atoms we see simulation speeds of 115 ns/day and 123 ns/day using OpenMM and NAMD3 respectively. Therefore, a TIES calculation with this system using

[1] https://ucl-ccs.github.io/TIES_MD/

[2] https://ccs-ties.org/[29]

13 alchemical windows and 5 replica simulations per window takes around 70 minutes of wall time using 65 V100s. In this work NAMD2 is run on the CPU platform and using 96 Xenon Skylake cores the simulation speed is 26 ns/day. Therefore, using NAMD2 and 6240 cores for one TIES calculation, again with 13 windows and 5 replicas, takes around 5-6 hours of wall time. In OpenMM the calculation of potentials and gradients, required for FEP and TI analysis, can be performed concurrently with the simulation, this creates an overhead of around 10% in TIES MD. The speed of OpenMM without this overhead is therefore 127 ns/day. For our NAMD calculations either the potential or gradient can be saved with the simulation but not both. Therefore the TI and FEP results cannot be collected concurrently and a post-processing step is needed to extract the FEP result from the NAMD trajectories. This post-processing generally takes 10-20 minutes for one TIES calculation. More details of the alchemical protocol will follow here; for the precise details of the MD package settings and performance refer to sections 6-8 of the Supplementary Information.

All the methods in this work use the same dual topology input systems. These systems model five proteins and 54 ligand transformations. The models are taken from the previous work of Bhati *et al.*[25] and details of their preparation are provided in that paper. In the SI of this paper we provide all these parameterised systems and note here that the AMBER ff99SB-ILDN[44] force field was used for protein parameters and the ligand parameters were produced using the general AMBER force field (GAFF)[45]. Our alchemical protocol involves collecting sampling from 13 intermediate alchemical states. This entails running an energy minimization followed by 2 ns of equilibration. After running pre-production on each state, 4 ns of NPT production sampling is run. In each state an ensemble of five simulations is performed for each simulation leg to calculate one ΔG value. From the production sampling the potential and gradient, $\frac{\partial u(\lambda,x)}{\partial \lambda}$, are calculated every 4 ps.

For the FEP estimator based results presented in this work, each one of the replicas in the ensemble of five simulated allowed for the calculation of one ΔG by applying equation 3 to the potentials sampled from the simulation. The five resulting values of ΔG are then bootstrapped to calculate a mean and standard error of the mean (SEM). For the TI results we apply the TIES protocol as it has been used in previous work[25]. The defining characteristic of TIES is the use of an ensemble of simulations in each alchemical state to control the aleatoric errors inherent to MD simulations. In every one of the total 13 alchemical states an ensemble of 5 simulation is performed, each of which yields a time series of $\frac{\partial u(\lambda,x)}{\partial \lambda}$ that can be averaged to give $\left\langle \frac{\partial u(\lambda,x)}{\partial \lambda} \right\rangle_\lambda$. An ensemble of five such values is then bootstrapped to calculate the mean which is used as the final value in equation 2. Each bootstrapping provides an estimate in the uncertainty, as a SEM, of the gradient in each alchemical window, $\sigma^2(\lambda)$, which is propagated as follows to give a total estimate of the uncertainty in each ΔG calculation

$$\sigma^2_{solvent/complex} = \sum_\lambda \sigma^2(\lambda)\Delta\lambda^2. \tag{6}$$

Here $\sigma^2_{solvent/complex}$ is the variance in one thermodynamic leg of the simulation and $\Delta\lambda$ is the difference between the value of λ between adjacent windows. The error from complex and solvent legs is combined in quadrature for both TIES and FEP methods to calculate the final uncertainty on the binding free energy.

4 OpenMM Alchemical Protocol

OpenMM does not offer any inbuilt alchemical methods and as such there exist a number of programs which extend OpenMM, allowing systems to be manipulated alchemically and perform alchemical calculations. One such program, used in this work, is OpenMMTools 0.19.0[46]. OpenMMTools can take as input a standard OpenMM system, defined with some potentials, and transforms this system into an alchemical one, where the potentials are controlled by the λ parameter. The scaling of (LJ) interactions was performed with a soft-core potential using the functional form of equation 13 presented in the work of Pham *et al.*[47] with parameters: $\alpha = 0.5$, $a = 1$, $b = 1$ and $c = 6$, the default parameters used by

OpenMMTools. Electrostatic interactions are scaled linearly without a soft-core potential. The λ schedule used in the OpenMM calculations was a two-step procedure which completely annihilated all electrostatic interactions of outgoing alchemical moieties before scaling down the LJ interaction, and completely created all LJ interactions of incoming moieties before turning on any electrostatic interactions. Annihilation was used in the OpenMM method as this is the methodology supported by OpenMMTools when calculating the electrostatics with the particle-mesh Ewald method, which was used for all simulations in this work. In this context, annihilation means that when a chemical moiety is “turned off” both inter and intra molecular interactions are extinguished.

5 NAMD Alchemical Protocol

Two versions of NAMD2/3 are used in this work, for CPU and GPU calculations. The alchemical protocol used is the same and so they will be discussed here jointly. The NAMD method uses a soft-core potential to decouple the LJ interactions. This soft-core potential can be expressed in the same form as the OpenMM soft-core using parameters $\alpha = 0.5$, $a = 1$, $b = 1$ and $c = 2$ these are the default parameters used by NAMD. Electrostatic interactions are decoupled linearly without a soft-core potential. The λ schedule used in the NAMD calculations was a one-step procedure where LJ and electrostatic potentials are scaled simultaneously. Decoupling was used in the NAMD method as this is the method invoked in our original NAMD2 based study of these systems[25]. In this context, decoupling means that when a chemical moiety is “turned off” only the inter molecular interactions are removed.

Whilst this procedure describes the simulation protocol accurately, one *caveat* must be added in the case of the NAMD2 results. The results presented here for NAMD2 are from the work of Bhati *et al*[25]. In this previous work the gradients, used in equation 2, are collected at intervals of 4 *ps* but the trajectories are saved at intervals of 10 *ps*. Therefore, the post-processing of FEP results can only be calculated at intervals of 10 *ps*. This affects the comparison of TI and FEP results and we address the matter as it arises in the analysis of the results.

6 Results

In this section, we present the wide range of results obtained in this study, covering comparisons between MD packages and free energy protocols, ensembles verses one-off simulations and the free energy distributions found.

6.1 Comparing Main Protocols

In the present work we study 54 ligand transformations in the protein targets MCL1, PTP1B, TYK2, CDK2 and thrombin. Here we present the results of these calculations comparing accuracy and precision across the MD packages and free energy estimators. Table 1 presents a comparison of the accuracy and precision of all methods compared to experiment.

In table 1 it can be seen that the results across all MD packages and estimators agreed well with one another. The only case for which a statistically significant difference can be observed is the PTP1B target. Comparing NAMD3 FEP and OpenMM FEP in the PTP1B case the values of MUE, MSE and RMSD differ by 0.20(0.21), 0.25(0.19) and 0.24(0.19) $kcal/mol$ respectively. Comparing NAMD2 FEP and OpenMM FEP, the MSE and RMSD differ by 0.24(0.20) and 0.22(0.21) $kcal/mol$ respectively. To investigate the PTP1B result further plots are presented in figure 1 for the FEP results in the case of PTP1B compared to experiment.

From figure 1 it can be seen that there are some statistically significant differences for the $\Delta\Delta Gs$ calculated with each method. Note that in figure 1 there are no error bars on the x-axis; this is because no errors are reported with the experimental results[25]. These differences in the PTP1B case should not detract from the overall excellent agreement between all

Table 1: Statistical properties calculated for all protein targets and alchemical methods. Properties are calculated with comparison to experimental data. Properties and 95% confidence intervals, provided in parentheses, are calculated with bootstrapping.

protein	**property**	**NAMD3 TI**	**NAMD3 FEP**	**NAMD2 TI**	**NAMD2 FEP**	**OpenMM TI**	**OpenMM FEP**
PTP1B	MUE	0.45(0.17)	0.36(0.11)	0.48(0.22)	0.36(0.16)	0.61(0.19)	0.60(0.18)
	MSE	0.28(0.15)	0.17(0.08)	0.35(0.18)	0.18(0.10)	0.46(0.20)	0.42(0.17)
	RMSD	0.53(0.16)	0.41(0.11)	0.59(0.19)	0.43(0.14)	0.68(0.17)	0.65(0.15)
	Pearson's	0.69(0.40)	0.81(0.33)	0.68(0.94)	0.83(0.38)	0.36(0.81)	0.48(0.83)
	slope	0.75(0.50)	0.93(0.61)	0.65(0.47)	0.80(0.48)	0.70(2.16)	0.96(1.52)
	intercept	0.16(0.62)	0.00(0.45)	0.13(0.73)	0.02(0.41)	0.31(1.19)	0.23(1.09)
CDK2	MUE	0.97(0.30)	0.96(0.30)	0.82(0.32)	0.76(0.38)	0.98(0.37)	0.98(0.40)
	MSE	1.03(0.42)	1.02(0.41)	0.83(0.43)	0.78(0.46)	1.41(0.86)	1.48(0.95)
	RMSD	1.02(0.25)	1.01(0.24)	0.91(0.26)	0.89(0.30)	1.19(0.45)	1.22(0.50)
	Pearson's	0.88(0.31)	0.88(0.32)	0.85(0.35)	0.83(0.58)	0.89(0.20)	0.88(0.18)
	slope	0.52(0.21)	0.52(0.21)	0.56(0.25)	0.57(0.28)	0.46(0.20)	0.45(0.20)
	intercept	-0.05(0.78)	-0.03(0.77)	-0.01(0.85)	0.01(0.77)	0.18(0.45)	0.17(0.47)
MCL1	MUE	1.35(0.36)	1.38(0.34)	1.17(0.30)	1.15(0.32)	0.98(0.34)	0.97(0.33)
	MSE	2.38(0.93)	2.48(0.93)	1.83(0.77)	1.86(0.78)	1.82(1.11)	1.80(1.13)
	RMSD	1.54(0.34)	1.57(0.33)	1.35(0.33)	1.37(0.32)	1.35(0.49)	1.34(0.52)
	Pearson's	0.81(0.22)	0.81(0.22)	0.81(0.22)	0.82(0.24)	0.74(0.49)	0.72(0.48)
	slope	0.52(0.20)	0.51(0.17)	0.56(0.18)	0.57(0.16)	0.70(0.41)	0.68(0.40)
	intercept	-0.11(0.46)	-0.11(0.47)	-0.05(0.48)	-0.11(0.47)	-0.42(0.40)	-0.34(0.41)
TYK2	MUE	0.67(0.20)	0.68(0.21)	0.42(0.20)	0.38(0.18)	0.62(0.20)	0.63(0.18)
	MSE	0.57(0.25)	0.58(0.24)	0.31(0.20)	0.27(0.16)	0.55(0.27)	0.53(0.24)
	RMSD	0.76(0.18)	0.76(0.18)	0.56(0.21)	0.52(0.19)	0.74(0.21)	0.73(0.19)
	Pearson's	0.90(0.15)	0.89(0.18)	0.94(0.10)	0.95(0.10)	0.89(0.18)	0.89(0.18)
	slope	0.94(0.29)	0.93(0.29)	1.12(0.28)	1.11(0.28)	1.05(0.35)	1.03(0.33)
	intercept	0.24(0.61)	0.22(0.58)	0.15(0.45)	0.11(0.45)	0.11(0.58)	0.12(0.54)
Thrombin	MUE	0.76(0.33)	0.77(0.30)	0.63(0.20)	0.68(0.25)	0.85(0.20)	0.82(0.20)
	MSE	0.94(0.55)	0.92(0.57)	0.49(0.18)	0.60(0.26)	0.87(0.34)	0.81(0.33)
	RMSD	0.97(0.32)	0.96(0.34)	0.7(0.15)	0.78(0.19)	0.93(0.21)	0.90(0.22)
	Pearson's	0.90(0.22)	0.92(0.13)	0.92(0.13)	0.92(0.17)	0.89(0.32)	0.89(0.28)
	slope	0.48(0.11)	0.49(0.10)	0.59(0.10)	0.55(0.10)	0.49(0.10)	0.50(0.10)
	intercept	0.03(0.26)	0.04(0.23)	-0.02(0.33)	0.04(0.27)	0.14(0.27)	0.10(0.28)

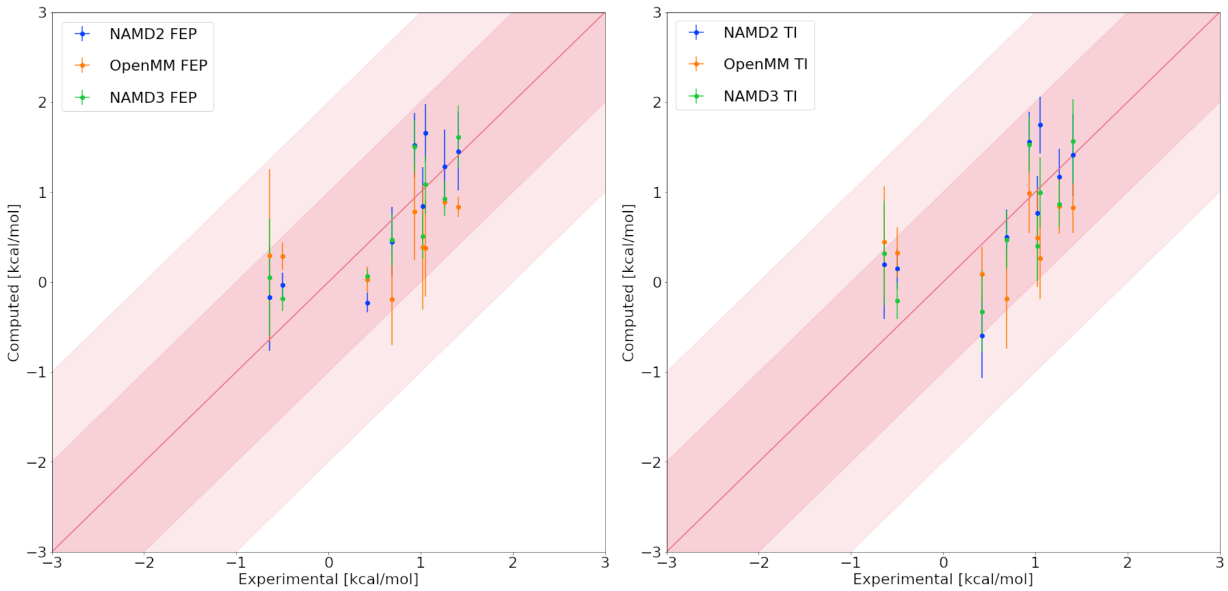


Figure 1: Computed vs experimental $\Delta\Delta G$ values for the PTP1B target using TI and FEP analysis methods. Computed $\Delta\Delta G$ and errors are average and standard deviations from bootstrapping an ensemble of 5 replicas. Dark shaded region spans ± 1 $kcal/mol$; lighter region spans ± 2 $kcal/mol$.

other cases and methods; in fact some difference in the results from different MD packages should be expected due to the unavoidable differences in implementation detailed in the methods section. The difference in individual $\Delta\Delta G$ calculated with different MD packages and free energy estimators is shown in table 2 where the difference between methods is around 0.5 $kcal/mol$. Due to the number of differences between methods, highlighted in previous sections, it is not possible to comment on what precisely causes any particular difference here. Despite some differences for individual $\Delta\Delta G$ calculations in MD packages, overall the results are well reproduced which can also be seen from the properties calculated in table 2 where the rank order coefficients suggest a strong correlation between all methods with the lowest Spearman's, Pearson's and Kendall's correlation coefficient between two packages being 0.91(0.04), 0.89(0.06) and 0.74(0.08) respectively.

A key result from table 1 is that there is no statistically significant difference between the calculated properties for TI and FEP results in all cases. In order to make the comparison between FEP and TI more rigorously, the calculated $\Delta\Delta Gs$ for each ligand transformations

Table 2: Statistical properties measuring the agreement between $\Delta\Delta Gs$ calculated from different MD packages in the TI and FEP cases. Properties and 95% confidence intervals, provided in parentheses, are calculated with bootstrapping.

Estimator	**property**	**OpenMM/NAMD2**	**OpenMM/NAMD3**	**NAMD2/NAMD3**
TI	MUE	0.51(0.13)	0.58(0.18)	0.48(0.17)
	MSE	0.48(0.23)	0.74(0.58)	0.57(0.43)
	RMSD	0.69(0.15)	0.86(0.28)	0.75(0.26)
	Spearman's	0.92(0.04)	0.91(0.04)	0.92(0.05)
	Pearson's	0.91(0.04)	0.89(0.06)	0.92(0.05)
	Kendall's Tau	0.77(0.08)	0.74(0.08)	0.81(0.08)
	slope	0.86(0.15)	1.06(0.15)	1.03(0.13)
	intercept	0.04(0.19)	-0.03(0.22)	-0.03(0.22)
FEP	MUE	0.50(0.11)	0.41(0.13)	0.50(0.17)
	MSE	0.42(0.19)	0.37(0.30)	0.62(0.46)
	RMSD	0.65(0.14)	0.61(0.20)	0.79(0.25)
	Spearman's	0.95(0.02)	0.95(0.03)	0.94(0.03)
	Pearson's	0.93(0.03)	0.95(0.03)	0.91(0.05)
	Kendall's Tau	0.80(0.06)	0.84(0.06)	0.79(0.07)
	slope	0.84(0.13)	1.03(0.10)	1.10(0.16)
	intercept	-0.01(0.18)	-0.05(0.16)	0.01(0.20)

are compared individually. From this comparison five transformations are identified as having significantly different TI and FEP results. All differences are found for the thrombin target when using the NAMD methods. The OpenMM implementation had no statistically significant differences for any protein targets. In the NAMD2 case the significantly different transformations are l1-l8 and l2-l5 and in the NAMD3 case these are l8-l1, l1-l4 and l2-l5. These transformations are named in the work of Bhati *et al.*[25] and figure 2 shows these ligand transformations explicitly. We remark here that, based on the similarity in differences in NAMD2 and NAMD3, the previously mentioned *caveat* in the case of NAMD2 regarding the lower FEP sampling rate does not have a significant impact on the difference in TI and FEP results.

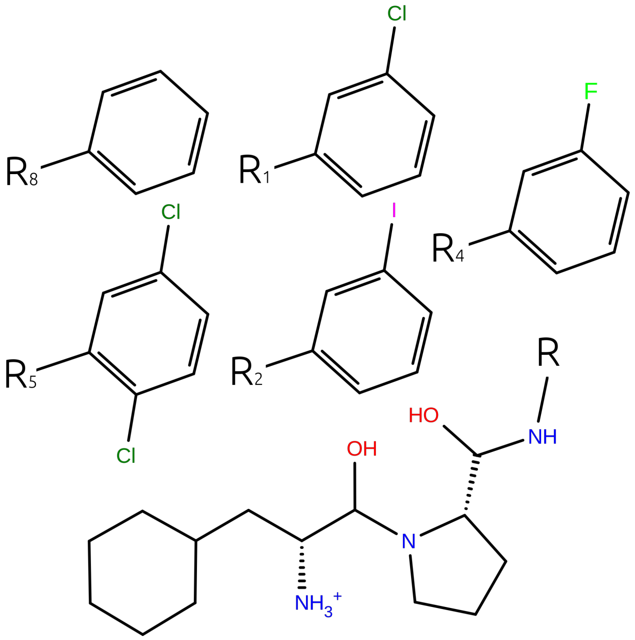


Figure 2: Labeled ligand transformations for which TI and FEP yield different result in NAMD TI and FEP methods. Moieties labelled R_x are substituted onto the common substructure at the position denoted by R; e.g. swapping R_1 and R_8 is the ligand transformation l1-l8.

6.2 Soft-core Potentials in TI Calculations

All the transformation in figure 2 feature the transformation of one phenyl group to another at the same location. From this similarity it might be concluded that something specific about the ligands causes the difference in TI and FEP results. However, we note that many results for the thrombin target feature similar transformations yet exhibit no significant differences.

Without a definitive relation to the specifics of the transformation, the cause of this difference is instead attributed to the behaviour of $\left\langle \frac{\partial u(\lambda,x)}{\partial \lambda} \right\rangle_\lambda$ at the end states for these transformations in the NAMD cases. This can be seen by plotting this gradient for the complex leg of the simulation across all states for the NAMD3 l2-l5 case in figure 3(a). In figure 3(a) we observed a rapid change for the gradient of the potential with respect to the lambda parameter which controls the LJ interactions of the disappearing alchemical region. Rapid changes such as this may result in poor accuracy for numerical integration and, without due care, this is known to be a weakness of the TI method[47]. This rapid change of the gradient is characteristic of all the transformations where we observe differences in the TI and FEP result. Moreover, these rapid changes are lessened or do not exist in the OpenMM case, explaining why no differences are observed.

In this work the key difference between OpenMM and NAMD methodologies, which pertained to the LJ interactions, lies in the parameters employed in the soft-core potential. The OpenMM method used c=6 whilst NAMD used c=2, so as such the OpenMM potential is softer. To test if a softer potential in NAMD can alleviate the difference in the TI and FEP calculations, the problematic transformations are repeated using NAMD3 with a soft-core potential with parameters $\alpha = 1.0$, $a = 1$, $b = 1$ and $c = 2$. Notice that α is modified here because c cannot be set by the user in NAMD. Table 3 shows the resulting $\Delta\Delta G$ values for the repeated calculation and the new differences with the equivalent FEP calculation. The results in table 3 show that there are no remaining significant differences in the TI and FEP result for these transformations. Additionally it can be seen in figure 3(b) that the gradient no longer features a rapid change in the final state. It should be noted that the choice of $\alpha = 1.0$ is not a good one and causes an increase in the variance in the final result. Other choices of soft-core parameters should be considered in general[40].

Table 3: Difference between FEP and TI result for three transformation in thrombin target re-run with NAMD3 with different values of the soft-core α parameter.

Transformation	α **= 0.5**	α **= 1**
l8-l1	-0.42(0.21)	-0.01(0.22)
l1-l4	0.46(0.15)	-0.22(0.48)
l2-l5	0.64(0.13)	-0.18(0.26)

6.3 Overlap Between Alchemical States in FEP Calculations

Another difference in the TI and FEP results may stem from the perturbative nature of FEP. If the phase space overlap of alchemical states is small, then the FEP result may be unreliable. A quantitative measure of this overlap of states can be made with an overlap

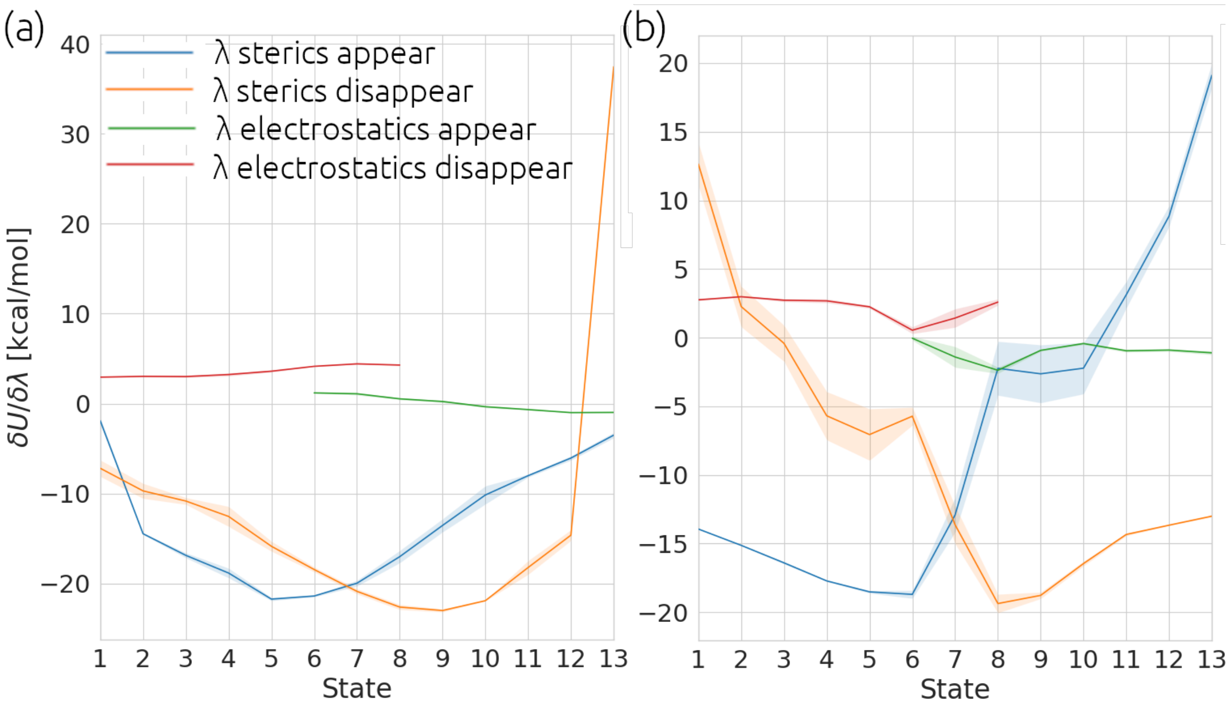
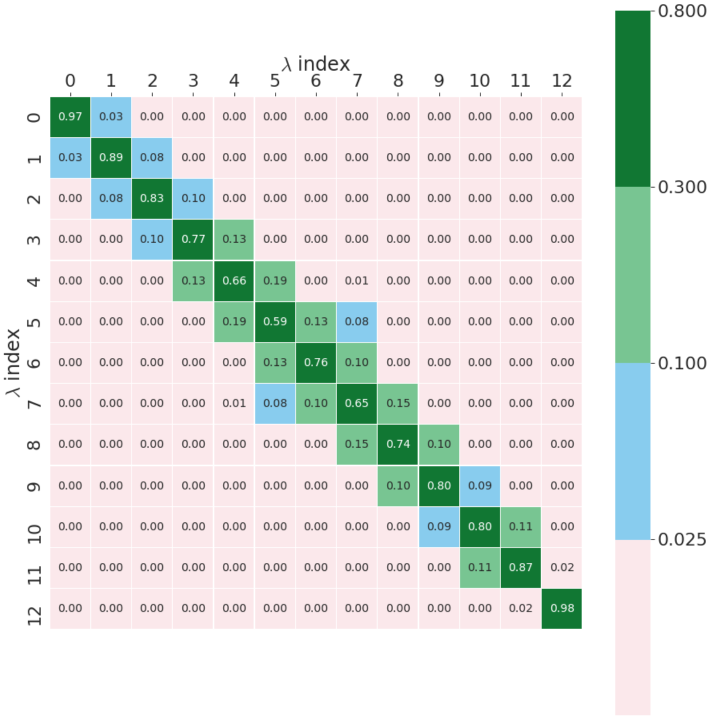


Figure 3: Gradients of potential with respect to λ parameters for transformation l2-l5 simulated with soft-core $\alpha = 0.5$ in sub figure (a) and $\alpha = 1.0$ in sub figure (b). Shaded regions show the mean ± SEM calculated from 5 replica calculations in each window.

matrix which was computed for all transformations and thermodynamic legs. The overlap matrix is described in detail elsewhere[43] but, briefly it is a matrix of rank $K \times K$, where K was previously defined in that work as the number of alchemical states. Each entry in the matrix is the probability that a sample from a given alchemical window λ_i could have been sampled from some other alchemical window λ_k. For reliable free energy calculations it has been proposed in previous work that the overlap matrices should be tri-diagonal with off-diagonal values greater than 0.03[41]. When the overlap matrices are averaged across replicas all but one of the FEP calculations performed in this work satisfied these conditions and this result is shown in figure 4. The simulation with the abnormal matrix is the complex leg of an OpenMM simulation for the MCL1 target. The abnormal transformation is named l12-l35; figure 5 shows this transformation explicitly. If the overlap matrices are not averaged over replicas there are more instances of results that do not reach the threshold of 0.03, these all occur for the complex leg of the MCL1 target simulations. Half of these low overlap cases are for the OpenMM protocol and 7 out of 8 of the cases are for transformations substantially similar to l12-l35 see; figure 6 for representative examples of such transformations. Without averaging over replicas the value of the overlap averaged over all instances failing to reach the 0.03 threshold is 0.02. If the same entries of the overlap are matrices are averaged over all replicas this value increases to 0.05. Despite lower overlap in some cases this does not manifest itself materially in any significant differences between the TI and FEP results for these transformations. To show this conclusively, table 4 exhibits the difference in ΔG results in the low overlap MCL1 case for the complex leg.

From the overall good agreement we find between the results calculated using the TI and FEP estimators we remark on the conclusion of previous studies[48], which compared Schrödinger's FEP+ to other TIES based alchemical methods revealing significant underestimation of the free energies when using FEP+. In this case, there are several sources of difference in the methodologies, including different force fields and FEP+ use of so-called "enhanced sampling". Since the present study finds good agreement between TI and FEP estimators it is clear that the remaining differences require further work to unravel these significant differences, although the proprietary nature of FEP+ does not make any such study straightforward.



λ index	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.97	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.03	0.89	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.08	0.83	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.10	0.77	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.13	0.66	0.19	0.00	0.01	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.19	0.59	0.13	0.08	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.13	0.76	0.10	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.01	0.08	0.10	0.65	0.15	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.74	0.10	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.80	0.09	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.80	0.11	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.87	0.02
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.98

Figure 4: Overlap matrix calculated for the OpenMM MCL1 l12-l35 complex simulation, averaged from five replicas.

6.4 Relative Free Energy Distributions

To examine the distribution of calculated binding free energies we selected the thrombin system and the OpenMM protocol to run larger ensembles of simulations. 48 simulations are run in all 13 lambda windows for 4 ns with all 11 ligands examined for the thrombin target. An analysis for these results is made one replica at a time and figure 7 shows examples of the distribution of the relative binding free energies which are found in the results. We plot these results with a calculation of the skewness and excess kurtosis. The skewness characterizes the symmetry of the distribution and kurtosis is related to the tails of the distribution, where higher values of the kurtosis indicates the presents of a significant number of outliers in the

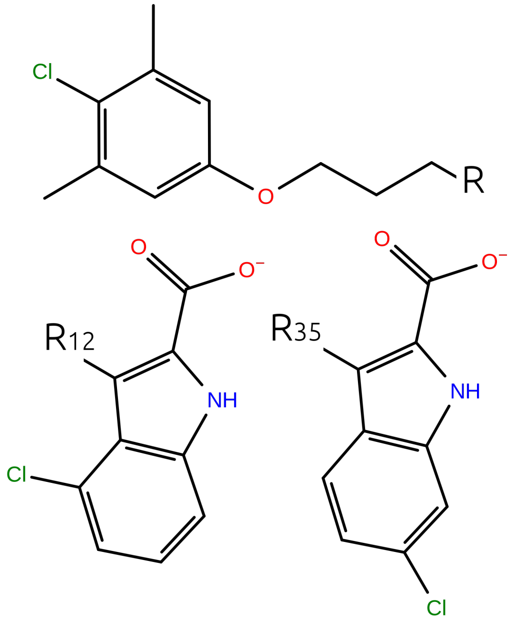


Figure 5: Substituted groups and common substructure for MCL1 transformations l12-l35. Moieties labelled R_x are substituted onto the common substructure at the position denoted by R; e.g. swapping R_{12} and R_{35} is the ligand transformation l12-l35.

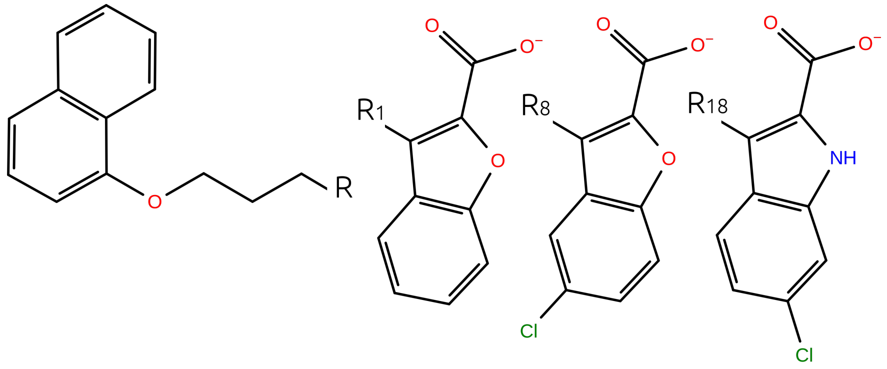


Figure 6: Substituted groups and common substructure for MCL1 transformations l1-l8, l8-l18. Moieties labelled R_x are substituted onto the common substructure at the position denoted by R; e.g. swapping R_1 and R_8 is the ligand transformation l1-l8.

Table 4: Difference in TI and FEP complex ΔG result for which overlap matrix exhibits indication of low overlap between adjacent states. Associated SEM computed by adding error on TI and FEP result in quadrature.

Method	**Transformation**	**TI-FEP**	**SEM**
NAMD2	l8-l18	0.09	0.98
	l1-l8	-0.27	1.15
	l16-l34	0.47	1.08
NAMD3	l1-l8	0.44	0.94
OpenMM	l8-l118	0.19	1.22
	l1-l8	-0.05	1.06
	l12-l35	-0.10	1.19
	l32-l38	-0.50	0.57

distribution. Here we report the “excess kurtosis” as the kurtosis-3. The excess kurtosis measures the deviation of the kurtosis with respect to the kurtosis one would expect for a Gaussian distribution. Figure 7 shows distributions of the binding free energy for two randomly selected ligand transformations; these distributions are approximately symmetric in terms of both skewness and excess kurtosis. If we examine all values of skewness and excess kurtosis as plotted in figure 8 as a function of the relative binding free energy it can be seen that although many results look approximately Gaussian there are distributions with significant skew and kurtosis. Overall these results imply the presence of non-normal distributions. This is consistent with previous work which has reported the same non-Gaussian distributions for relative binding free energy calculations[29].

The FEP result is shown in figure 8 alongside the TI result and these non-Gaussian results are mirrored in TI and FEP. Whilst this result seems to contradict previous work by Paliwal *et al.*[40] which found free energy distributions to be Gaussian, that work examined much simpler systems of small molecule hydration. In this work we consider “real world” molecular systems which are strongly influenced by anharmonic terms (e.g. van der Waals interactions), not performed in a homogeneous environment (e.g. in a protein) and exhibit more than one dominant conformational substrate[49]. The underlying nonlinearities in the dynamics are what accounts for both the presence of chaos and non-Gaussian statistics. In general a Gaussian distribution of free energy results can only be assumed for harmonic systems or transformations that can be approximated by linear response theory[50–52].

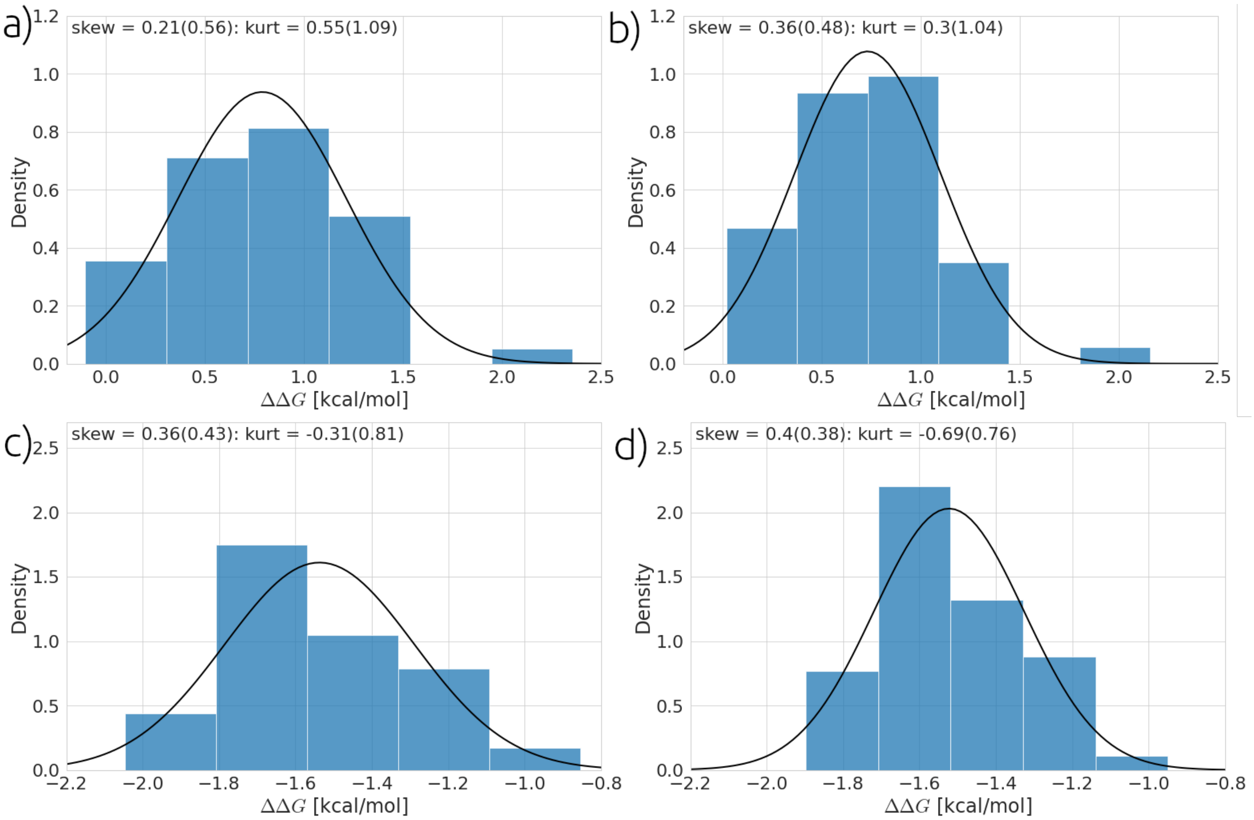


Figure 7: Panels a) and b) show the distribution of the relative binding free energies from for 48 simulations for the thrombin ligand transformation named l2-l5 with results estimated by TI and FEP respectively. Panels c) and d) show the distribution of the relative binding free energies from for 48 simulations for the thrombin ligand transformation named l1-l4 with results estimated by TI and FEP respectively. Parentheses provide 90% bootstrapped confidence intervals on calculation of skewness and excess kurtosis(kurt). The black line shows a Gaussian distribution with the same mean and σ as the plotted data.

6.5 Comparing Long and Large Ensemble Simulations

Previous work using this data set of input transformations and target proteins has demonstrated that an ensemble of 5 replica simulations using 13 alchemical windows with 4 ns of sampling per window provides a good trade off of computation cost against accuracy and precision. For completeness we re-examine this rule of thumb in the context of our work's larger set of free energy estimators and MD engines. To perform this comparison 6 ligand transformations are selected from the full set of 54 named in previous work[25] as l12-l35 and

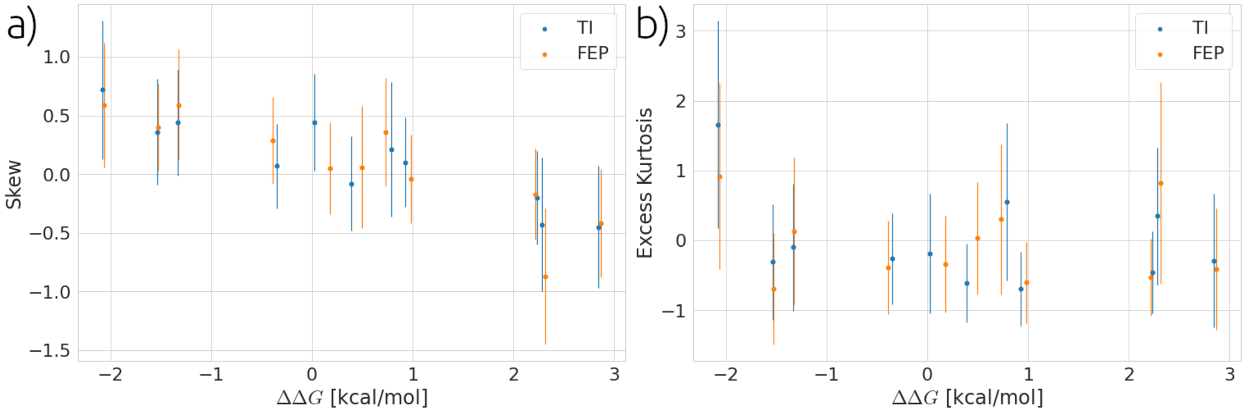


Figure 8: Panels a) and b) show the skewness and excess kurtosis for all 11 thrombin ligand transformations examined using both TI and FEP estimators. Error bars are plotted as 90% bootstrapped confidence intervals.

l16-l34 for the MCL1 target, l15-l10 and l16-l10 for the TYK2 target and l3-l23 and l13-l20 for the PTP1B target. These 6 transformations are then rerun using all estimators and engines with the same TIES methodology previously discussed but now modified in one of two ways. The first modification is to use 20 sets of 4 ns simulations instead of 5 sets of 4 ns per window, which we call large ensemble runs. The second modification is to use 5 sets of 40 ns runs per window, again instead of the typical 5 sets of 4 ns, which we call long runs.

In figure 9 we see a comparison of results collected using the long and large ensemble simulation protocols with one ligand transformation for the MCL1 target. What can be seen from the results in 9 is that even when using less production simulation the use of many independent and shorter simulations provides similar accuracy and better precision that using fewer and longer simulations. This is a repeated pattern: figure 9c shows that the error on the large ensemble runs is much lower than that of the long runs when averaged over all 6 transformations examined here. Table 5 compares the accuracy of large ensemble and long runs and shows that, indeed, averaged over all ligand transformations, the accuracy of these methods is similar despite using less overall simulation time in the large ensemble runs.

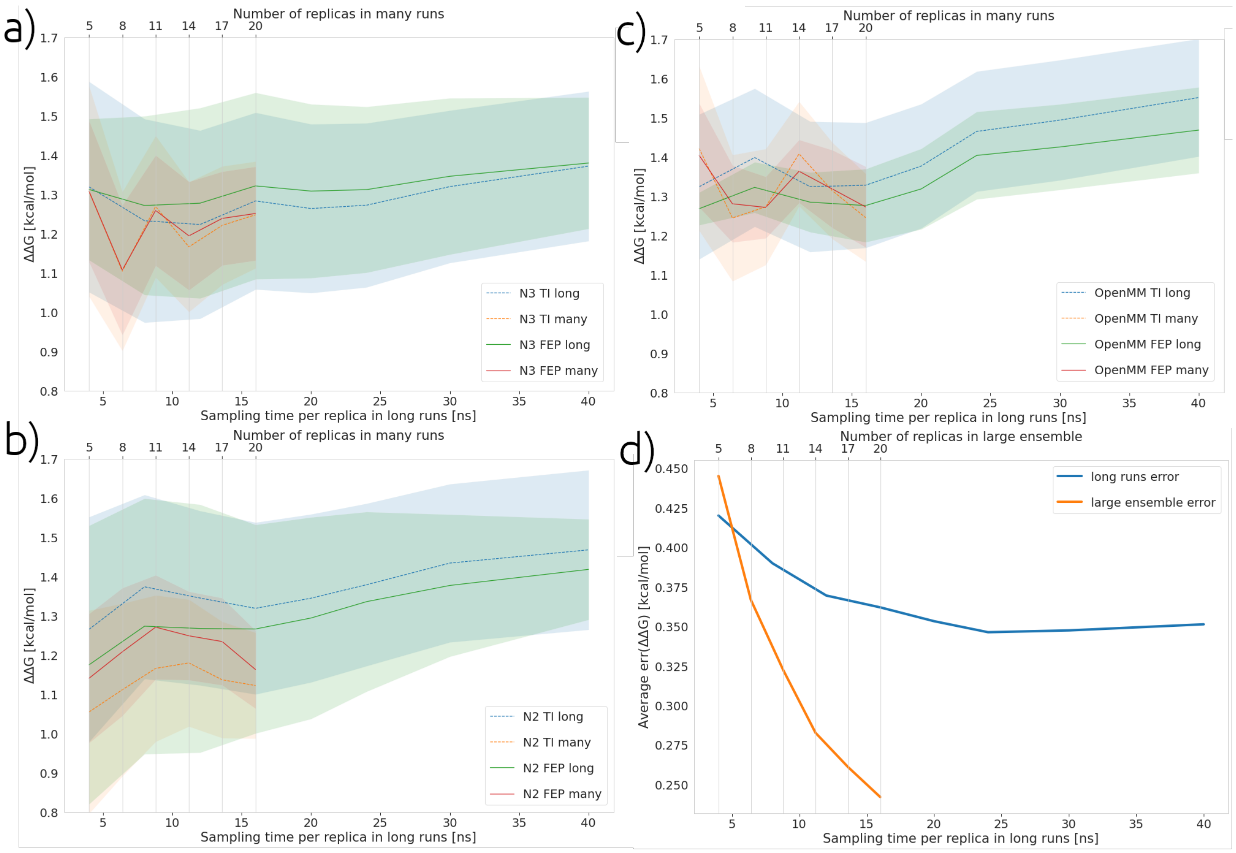


Figure 9: Calculated relative binding free energies for ligand transformation l15-l16 from the target TYK2 (experimental result for this ligand is 0.75 kcal/mol). This figure compares long and large ensemble simulation protocols. Panels a), b) and c) show the results acquired using NAMD3(N3), OpenMM and NAMD2(N2) respectively. Panel d) plots the average statistical uncertainty for all transformations, again comparing long and large ensemble simulation protocols. Shaded regions show the mean ± SEM calculated from 5 replicas.

Table 5: Comparing the accuracy of large ensemble and long run simulation methods using 20 replicas (large ensemble runs), 40 *ns* per replica (long runs) or the standard TIES protocols for all six ligand transformations studied. Properties and 95% confidence intervals, provided in parentheses, are calculated with bootstrapping.

Protocol	**Property**	**NAMD2 TI**	**NAMD2 FEP**	**OpenMM TI**	**OpenMM FEP**
large ensemble runs	MUE	0.63(0.78)	0.60(0.65)	0.60(0.58)	0.62(0.57)
	MSE	1.12(2.09)	0.83(1.44)	0.80(1.18)	0.77(1.18)
	RMSD	1.06(0.73)	0.91(0.60)	0.90(0.51)	0.88(0.51)
long runs	MUE	0.78(0.57)	0.64(0.48)	0.92(0.33)	0.84(0.29)
	MSE	1.02(1.32)	0.70(0.86)	1.02(0.65)	0.84(0.50)
	RMSD	1.01(0.52)	0.84(0.43)	1.01(0.28)	0.91(0.25)
standard TIES	MUE	0.55(0.41)	0.45(0.43)	0.80(0.62)	0.75(0.46)
	MSE	0.53(0.67)	0.43(0.71)	1.09(1.56)	0.84(1.02)
	RMSD	0.73(0.36)	0.65(0.41)	1.05(0.59)	0.92(0.45)

6.6 Calculating Uncertainty with One-off Simulations

The comparisons between different MD engines and free energy estimators, made in this work, could only be made meaningfully when the uncertainty in the binding free energy is accounted for correctly. The results from one-off simulations are not reproducible and so only with the proper application of ensemble simulation could such good agreement between the MD engines and free energy estimations compared in this work be found. The application of multiple independent simulations was critical for our error control; similar ideas are found elsewhere in the literature[31,41,53,54]. If only one-off simulations are performed errors are consistently underestimated in these calculations. Figure 10 shows this explicitly by comparing the SEM computed by MBAR averaged over 5 replicas to the “TIES-like” error calculated by computing the SEM of the bootstrapping results from 5 replicas. It can be seen that, for the ligand simulations in figure 10b, some systems (TYK2, CDK2 and thrombin) have errors correctly estimated by MBAR but for, PTP1B and MCl1, MBAR consistently underestimates the error. This underestimation is only exacerbated in complex simulation, Figure 10a, where all systems have their error underestimated by MBAR. This is most likely due to the greater relevance of “rare events” in the complex simulation. Similar findings by Rizzi *et al.* have concluded “Nevertheless, when sampling is governed by rare events and systematically misses relevant areas of conformational space, data from a single trajectory simply cannot contain sufficient information to estimate the uncertainty accurately”[39]. It has often been argued that the time series of potentials fed to MBAR should be decorrelated to ensure reliable error estimation[39]. Decorrelation of the time series of potentials, in this case, does not change any of the conclusions. For completeness, we provide an equivalent version of figure 10 using decorrelated data in the SI (see figure S1) which demonstrates this conclusively.

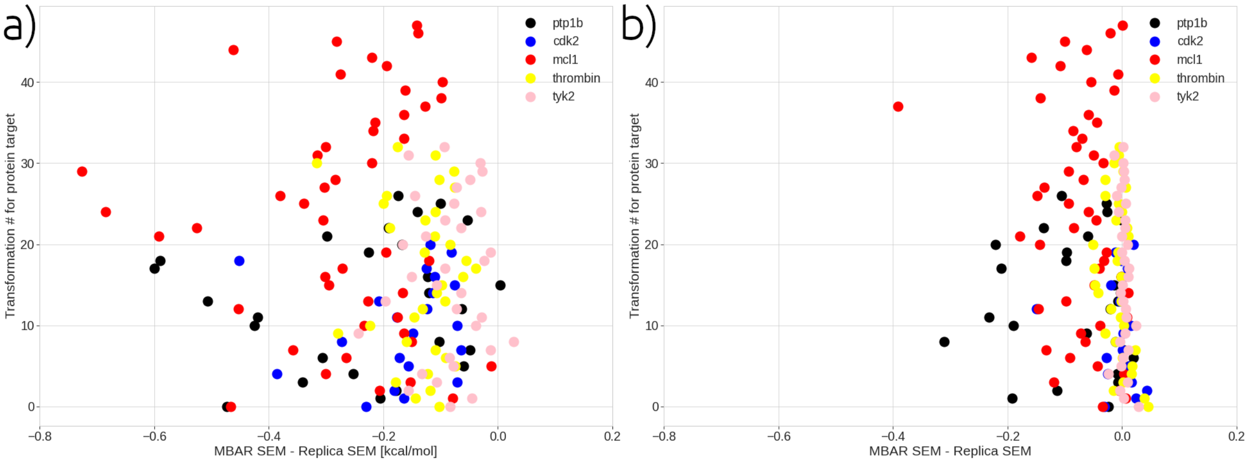


Figure 10: A comparison of the SEM estimated by MBAR from one replica and then averaged over 5 replicas, compared to a "TIES-like" error calculated by computing SEM of the bootstrapping result of 5 replicas. Panels a) and b) show the results for the ligand-protein and ligand only simulations respectively. The x-axis denotes an index assigned to each ligand transformation. This index runs from zero to the total number of transformations minus one, within each protein target across all engines.

7 Conclusions

In this work 54 ligand transformations for 5 diverse protein targets: MCL1, PTP1B, TYK2, CDK2 and thrombin have been examined and relative binding free energy calculations performed using three MD packages NAMD2, NAMD3 and OpenMM. Analysis of these calculations was made using two free energy estimators: TI and FEP. Effort was made to match the alchemical methodologies between MD packages as closely as possible. However, differences were observed such as diverse soft-core parameters, λ schedules, methods for calculating the TI gradient and the use of either decoupling or annihilating methods to turn off the alchemical regions. Despite these differences, the results show good agreement between molecular dynamics packages with an average mean unsigned error between packages of 0.5 $kcal/mol$. The correlation between packages was very good with the lowest Spearman's, Pearson's and Kendall's tau correlation coefficient between pairs of packages being 0.91, 0.89 and 0.74 respectively.

Comparing the TI and FEP estimators, two potential sources of error were investigated in detail. In the case of TI, errors can accumulate in the numerical integration if the potential has large curvature with respect to the alchemical parameters. This resulted in significant differences in the $\Delta\Delta G$ values estimated with TI and FEP. We saw this error manifest in NAMD calculations when examining the thrombin target. For the affected cases, rapid changes of the gradient in the initial and final alchemical states of the calculation were related to the soft-core parameters. Both OpenMM, which used a softer soft-core potential, and NAMD, when rerun with a softer soft-core potential, exhibited no significant difference in the TI and FEP results. The second source of differences in TI and FEP results investigated here stemmed from the overlap in energy distributions of neighbouring alchemical states. FEP based analysis techniques may be unreliable if neighbouring states do not have overlap in the sampled energy distributions. From the 324 relative binding free energy calculations performed in this work, one was identified as having a low overlap: this was l12-l35 in the MCL1 target. However, this low overlap was not found to translate into a difference in the TI and FEP results. Low overlap was found more frequently if analysis was made for "one-off" simulations but averaging over many replicas eliminated this in all but the one MCL1 case.

With the exception of the TI cases with rapid changes of the gradient in the end states, both TI and FEP methods achieve comparable accuracy and precision for the systems studied in this work and as such neither TI or MBAR is highlighted as preferred. In the occasional TI case which provided problems with the end states, issues were treated easily by choosing appropriate soft-core parameters. What is clear is the benefit of using both TI and FEP results in tandem to check the results of the other. Whilst using many MD packages to check the reproducibility of results incurs significantly more cost and may not always be practical, the use of two or more free energy estimators incurs little additional cost and, as shown in this work, can aid in the identification and diagnosis of alchemical protocol, specific issues causing poor accuracy or precision in the results[43].

Non-Gaussian distributions are observed for the free energies calculated using both TI and MBAR estimators. This assessment of the distribution was made for all 11 transformations of the thrombin target using the OpenMM protocol developed in this work with an ensemble of

48 simulations. The findings of both skewness and kurtosis in the distributions of calculated free energies are consistent with previous work[29] using NAMD2. It was also found, for 6 out of 54 transformations examined with the modified TIES protocols across all MD engines and estimators, that the use of large ensembles of shorter simulations as compared to smaller ensembles of longer simulations could yield comparable accuracy and improved precision. This was true even when using less overall production simulation time in the large ensemble cases.

Finally, the use of "one-off" simulations in the estimation of errors for protein-ligand RBFE calculations was considered. We compared the errors estimated by MBAR from a single simulation to the bootstrapped error from an ensemble of simulations and conclusively demonstrated that single simulations consistently underestimate errors. This is particularly true in the case of the ligand-protein complex simulations.

From these conclusions we summarise that relative binding free energy calculations performed in different MD packages are reproducible and moreover TI and MBAR estimators also produce reproducible results. This reproducibility can only be measured reliably with appropriate control of uncertainty and it is clear that "one-off" simulations do not provide reliable estimates of uncertainty.

8 Acknowledgements

The authors acknowledge funding support from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement 823712 (CompBioMed2, compbiomed.eu) and NSF Award (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1713749, Award No. NSF 1713749); EPSRC for the UKCOMES HighEnd Computing Consortium (grant no. EP/R029598/1) and for the Software Environment for Actionable and VVUQ-evaluated Exascale Applications (SEAVEA) (grant no. EP/W007711/1); The MRC Medical Bioinformatics project (MR/L016311/1), the EU H2020 projects ComPat (https://www.

compatproject.eu/, Grant No. 671564), and funding from the UCL Provost. We acknowledge the Gauss Centre for Supercomputing for providing computing time on the GCS supercomputer SuperMUC-NG (https://doku.lrz.de/display/PUBLIC/SuperMUC-NG) at Leibniz Supercomputing Centre under project COVID-19-SNG1 and the very able assistance of its scientific support staff. Furthermore, the authors made use of the following OLCF and ALCF supercomputers: Summit, Andes, ThetaGPU, and Theta thanks to a 2021 DOE INCITE award 'COMPBIO'. The authors are grateful to Mateusz Bieniek for numerous helpful and insightful discussions which contributed to materially improving this work.

9 Supporting Information

The Supporting Information for this work contains all individual results for relative binding free energy calculations performed as well as the input used to generate these results https://zenodo.org/record/5767275#.YbCZEXX7SV4. The results provided in the SI come from the main TIES protocol (table S1-S6) and the long and large ensemble TIES protocols table S7. Additionally we present an equivalent version of figure 10 plotted using decorrelated data (figure S1). We provide information on the long time simulations performed in this work with NAMD3 and highlight some statistically significant results we observe in figures S2-S3. These figures are complemented by an assessment of the accuracy of NAMD3 calculations across different simulation protocols in Table S8. Table S9 provides the PDB codes for the proteins used as input to this work. Table S10 presents more detailed information for the performance of MD engines at different system sizes. Finally, we provide a full list of settings used in the different MD engines.

References

(1) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.;

Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(2) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E. Picosecond to millisecond structural dynamics in human ubiquitin. *J. Phys. Chem. B* **2016**, *120*, 8313–8320.

(3) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

(4) Salomon-Ferrer, R.; Gotz, 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*, 3878–3888.

(5) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

(6) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending force field coverage for drug-like small molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1863–1874.

(7) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmer-

ling, C. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(8) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr., A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(9) The Open Force Field Initiative. https://openforcefield.org/, Accessed: 2021-09-18.

(10) Qiu, Y.; Nerenberg, P. S.; Head-Gordon, T.; Wang, L.-P. Systematic optimization of water models using liquid/vapor surface tension data. *J. Phys. Chem. B* **2019**, *123*, 7061–7073.

(11) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **2016**, *47*, 20–33.

(12) Steinbrecher, T. B.; Dahlgren, M.; Cappel, D.; Lin, T.; Wang, L.; Krilov, G.; Abel, R.; Friesner, R.; Sherman, W. Accurate binding free energy predictions in fragment optimization. *J. Chem. Inf. Model.* **2015**, *55*, 2411–2420.

(13) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic acids research* **2018**, *46*, D1074–D1082.

(14) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of generative autoencoder in *de novo* molecular design. *Mol. Inform.* **2018**, *37*, 1700123.

(15) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: an AI tool for *de novo* drug design. *J. Chem. Inf. Model.* **2020**, *60*, 5918–5922.

(16) Bhati, A. P.; Wan, S.; Alfè, D.; Clyde, A. R.; Bode, M.; Tan, L.; Titov, M.; Merzky, A.; Turilli, M.; Jha, S.; Highfield, R. R.; Rocchia, W.; Scafuri, N.; Succi, S.; Kranzlmüller, D.; Mathias, G.; Wifling, D.; Donon, Y.; Di Meglio, A.; Vallecorsa, S.; Ma, H.; Trifan, A.; Ramanathan, A.; Brettin, T.; Partin, A.; Xia, F.; Duan, X.; Stevens, R.; Coveney, P. V. Pandemic drugs at pandemic speed: infrastructure for accelerating COVID-19 drug discovery with hybrid machine learning-and physics-based simulations on high-performance computers. *Interface Focus* **2021**, *11*, 20210018.

(17) Wade, A. D.; Huggins, D. J. Identification of optimal ligand growth vectors using an alchemical free-energy method. *J. Chem. Inf. Model.* **2020**, *60*, 5580–5594.

(18) Coveney, P. V.; Wan, S. On the calculation of equilibrium thermodynamic properties from molecular dynamics. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30236–30240.

(19) Genheden, S.; Ryde, U. How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.* **2010**, *31*, 837–846.

(20) Sadiq, S. K.; Wright, D. W.; Kenway, O. A.; Coveney, P. V. Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J. Chem. Inf. Model.* **2010**, *50*, 890–905.

(21) Wright, D. W.; Hall, B. A.; Kenway, O. A.; Jha, S.; Coveney, P. V. Computing clinically relevant binding free energies of HIV-1 protease inhibitors. *J. Chem. Theory Comput.* **2014**, *10*, 1228–1241.

(22) Wan, S.; Bhati, A. P.; Zasada, S. J.; Wall, I.; Green, D.; Bamborough, P.; Coveney, P. V. Rapid and reliable binding affinity prediction of bromodomain inhibitors: a computational study. *J. Chem. Theory Comput.* **2017**, *13*, 784–795.

(23) Wan, S.; Bhati, A. P.; Skerratt, S.; Omoto, K.; Shanmugasundaram, V.; Bagal, S. K.; Coveney, P. V. Evaluation and characterization of Trk kinase inhibitors for the treatment of pain: Reliable binding affinity predictions from theory and computation. *J. Chem. Inf. Model.* **2017**, *57*, 897–909.

(24) Lawrenz, M.; Baron, R.; McCammon, J. A. Independent-trajectories thermodynamic-integration free-energy changes for biomolecular systems: determinants of H5N1 avian influenza virus neuraminidase inhibition by peramivir. *J. Chem. Theory Comput.* **2009**, *5*, 1106–1116.

(25) Bhati, A. P.; Wan, S.; Wright, D. W.; Coveney, P. V. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theory Comput.* **2017**, *13*, 210–222.

(26) Bhati, A. P.; Wan, S.; Hu, Y.; Sherborne, B.; Coveney, P. V. Uncertainty quantification in alchemical free energy methods. *J. Chem. Theory Comput.* **2018**, *14*, 2867–2880.

(27) Vassaux, M.; Wan, S.; Edeling, W.; Coveney, P. V. Ensembles are required to handle aleatoric and parametric uncertainty in molecular dynamics simulation. *J. Chem. Theory Comput.* **2021**, *17*, 5187–5197.

(28) Wan, S.; Sinclair, R. C.; Coveney, P. V. Uncertainty quantification in classical molecular dynamics. *Philos. Trans. Royal Soc. A* **2021**, *379*, 20200082.

(29) Bieniek, M. K.; Bhati, A. P.; Wan, S.; Coveney, P. V. TIES 20: Relative Binding Free Energy with a Flexible Superimposition Algorithm and Partial Ring Morphing. *J. Chem. Theory Comput.* **2021**, *17*, 1250–1265.

(30) Chipot, C. Frontiers in free-energy calculations of biological systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 71–89.

(31) 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.

(32) Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **2016**, *7*, 207–218.

(33) Deng, Y.; Roux, B. Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

(34) Procacci, P.; Chelli, R. Statistical Mechanics of Ligand–Receptor Noncovalent Association, Revisited: Binding Site and Standard State Volumes in Modern Alchemical Theories. *J. Chem. Theory Comput.* **2017**, *13*, 1924–1933.

(35) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated protein structure and topology generation for alchemical perturbations. *J. Comput. Chem.* **2015**, *36*, 348–354.

(36) Wang, L.; Chambers, J.; Abel, R. In *Biomolecular Simulations: Methods and Protocols*; Bonomi, M., Camilloni, C., Eds.; Springer New York: New York, NY, 2019; pp 201–232.

(37) Woods, C. FESetup: Automating setup for alchemical free energy simulations. *J. Chem. Inf. Model.* **2015**,

(38) TIES Toolkit. https://www.ties-service.org, Accessed: 2021-09-16.

(39) Rizzi, A.; Jensen, T.; Slochower, D. R.; Aldeghi, M.; Gapsys, V.; Ntekoumes, D.; Bosisio, S.; Papadourakis, M.; Henriksen, N. M.; De Groot, B. L.; Cournia, Z.; Dickson, A.; Michel, J.; Gilson, M. K.; Shirts, M. R.; Mobley, D. L.; Chodera, J. D. The SAMPL6 SAMPLing challenge: Assessing the reliability and efficiency of binding free energy calculations. *J. Comput. Aided Mol. Des.* **2020**, *34*, 601–633.

(40) 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*, 4115–4134.

(41) Mey, A. S.; Allen, B.; Macdonald, H. E. B.; 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 preprint arXiv:2008.03067* **2020**,

(42) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(43) Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the analysis of free energy calculations. *J. Comput. Aided Mol. Des.* **2015**, *29*, 397–411.

(44) Wang, B.; Li, L.; Hurley, T. D.; Meroueh, S. O. Molecular recognition in a diverse set of protein–ligand interactions studied with molecular dynamics simulations and end-point free energy calculations. *J. Chem. Inf. Model.* **2013**, *53*, 2659–2670.

(45) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(46) Chodera, J.; Rizzi, A.; Naden, L.; Beauchamp, K.; Grinaway, P.; Fass, J.; Wade, A.; Rustenburg, B.; Ross, G. A.; Krämer, A.; Macdonald, H. B.; Rodríguez-Guerra, J.; dominicrufa,; Simmonett, A.; Swenson, D. W.; hb0402,; Henry, M.; Roet, S.; Silveira, A. Choderalab/OpenMMTools: 0.20.3 Bugfix Release. **2021**,

(47) Pham, T. T.; Shirts, M. R. Identifying low variance pathways for free energy calculations of molecular transformations in solution phase. *J. Chem. Phys.* **2011**, *135*, 034114.

(48) Wan, S.; Tresadern, G.; Pérez-Benito, L.; van Vlijmen, H.; Coveney, P. V. Accuracy

and precision of alchemical relative free-energy predictions with and without replica-exchange. *Adv. Theory Simul* **2020**, *3*, 1900195.

(49) 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.* **2018**, *15*, 1265–1277.

(50) König, G.; Brooks, B. R.; Thiel, W.; York, D. M. On the convergence of multi-scale free energy simulations. *Mol. Simul.* **2018**, *44*, 1062–1081.

(51) Shirts, M. R.; Pande, V. S. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.* **2005**, *122*, 144107.

(52) Hummer, G.; Pratt, L. R.; Garcia, A. E. Free energy of ionic hydration. *J. Phys. Chem* **1996**, *100*, 1206–1215.

(53) Gapsys, V.; Yildirim, A.; Aldeghi, M.; Khalak, Y.; van der Spoel, D.; de Groot, B. L. Accurate absolute free energies for ligand–protein binding based on non-equilibrium approaches. *Commun. Chem.* **2021**, *4*, 1–13.

(54) Baumann, H. M.; Gapsys, V.; de Groot, B. L.; Mobley, D. L. Challenges Encountered Applying Equilibrium and Nonequilibrium Binding Free Energy Calculations. *J. Phys. Chem. B* **2021**, *125*, 4241–4261.