**On the Issues Impacting Reproducibility of Alchemical Free Energy Calculations**

Miroslav Suruzhon, Marley L. Samways & Jonathan W. Essex

**Introduction**

Alchemical free energy (AFE) calculations are commonly considered to be the most rigorous approach to estimate ligand binding affinities in drug discovery.1 Since each AFE calculation typically involves multiple molecular dynamics (MD) simulations at intermediate states (λ windows), all limitations of conventional MD apply to AFE calculations, as well as the additional complications introduced by the free energy estimation. These include the need for enhanced sampling to improve precision and for improved force fields which will result in higher accuracy.

Compared to force field development and enhanced sampling, an often-overlooked topic in the field of molecular simulation is that of reproducibility. In this chapter, we will distinguish between repeatability, which is directly related to the precision problem, and reproducibility, which is the ability of two different scientific groups to produce “statistically similar” results given the same problem and methodology. However, since simulation studies rely on many implicit choices by the researcher, these are only partially described in practice, meaning that the vast majority of MD calculations are inherently difficult to reproduce. These decisions are not only method-specific but are also related to how the researcher chooses to represent the real-world physical system as a computational model and what the purpose of the simulation is. Therefore, we will here and henceforth extend the concept of reproducibility to the ability to obtain the same free energy value using two *different* protocols.

There are two main stages of a free energy calculation: setup and simulation. The former includes the following choices: the initial coordinates of the system (e.g. protein crystal structure, ligand binding mode and water placement); a suitable force field for each of the components of the system; amino acid and ligand protonation states; the addition of extra components to the system, such as salt ions. The second stage of the free energy calculation process involves choices related to the sampling method (e.g. sampling time, number of repeats, integrator, thermostat and barostat), the free energy methodology used and, finally, all decisions related to the interpolation of the intermediate states (e.g. alchemical protocol, softcore potential functional form and parameters).

Both of the aforementioned stages entail many potential sources of reproducibility issues, and in the following sections we will outline some of the main studies which investigate the robustness of relative AFE calculations with respect to these choices, as well as methods to remedy reproducibility problems. In some cases, such as comparing different force fields, there will be no obvious way to reconcile discrepancies in the free energy values obtained from different methods and in these cases comparison to experiment will be reported instead. It needs to be noted, however, that these comparisons always assume that the experimentally obtained free energies are sufficiently reliable and that this condition is not always met in practice.2 Most of the following considerations are not unique to AFE calculations and are relevant to all molecular simulations. This chapter will therefore also shed light on potential sources of reproducibility issues in other applications of MD simulations.

**Simulation Setup**

**Initial coordinates**

**Protein crystal structure**

Owing to the limitations of current computing power, ab initio prediction of the three-dimensional structure of a protein-ligand complex remains a computationally challenging problem and is not currently feasible in the context of drug discovery. As such, the choice of initial coordinates is crucial for ensuring the physical validity of the MD simulation and needs to be obtained in a reliable and reproducible way.

Although it is common practice to use a structure experimentally resolved by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy to provide the initial coordinates, problems arise when no such structure is available. In this case one must choose a somewhat unrelated complex to derive the initial coordinates. This can, for example, be achieved by homology modelling3. However, despite the introduction of automated homology modelling workflows4, these methods remain sensitive to choices in the computational protocol, making them difficult to reproduce5.

Another difficulty specific to AFE calculations is the choice of the initial structure of the intermediate λ windows. For an alchemical perturbation of two ligands $A\rightarrow B$, it is not clear which of these two complexes should be used to model the intermediate states, making this choice largely arbitrary and not reproducible. In addition, even though two free energy calculations with different initial coordinates and identical topology should eventually converge to the same value with infinite sampling, the protein initial coordinates can also influence most of the setup process, including initial protonation states and water placement (these are discussed in later sections).

The above considerations show that obtaining initial protein structures is a complex process which not only directly impacts the sampling quality, but can also affect the asymptotic free energy value. Since different researchers will generally have different approaches to choosing an initial set of coordinates, it is important to be aware of the extent to which this lack of reproducibility can impact the resulting free energy results in a practical setting.

The choice of the initial protein coordinates has been extensively investigated in a study by Suruzhon *et al.*6 By performing a range of perturbations across three different protein systems, the authors compared the free energy differences between eight different initial crystal structures using the same binding mode. It was found that at commonly used timescales ($\~$4 ns per λ window), the choice of initial crystal structure can result in significant free energy discrepancies – sometimes larger than 1 kcal/mol. These results are in accordance with a previous study by Pérez-Benito *et al.* which demonstrated that the free energy discrepancies which arise from using initial coordinates from either ligand A or ligand B can often surpass 1 kcal/mol, and even changing the sign of a free energy which is presumed to be sufficiently favorable in a certain direction.7

Even more challenging are proteins with slow conformational modes. In these cases, the initially resolved crystal structure is no longer sufficient for obtaining representative sampling, and long-timescale or enhanced sampling becomes essential for these systems. This was shown in the case of T4-lysozyme by Lim *et al.*, who reported a root-mean-square deviation (RMSD) of 4 kcal/mol between the results obtained from the open and closed conformations.8 In these cases temperature-based enhanced sampling is a useful tool in overcoming relevant high kinetic barriers and the authors reported a marked increase in consistency when protein replica exchange with solute tempering (pREST) was used alongside the calculation to reduce the RMSD to 0.57 kcal/mol.

**Ligand binding mode**

If the ligand under consideration has some degree of translational, rotational or conformational flexibility, it may demonstrate multiple binding modes with the protein – this is especially likely for smaller, fragment-like compounds, which may bind in multiple locations and/or orientations9. If the ligand is not able to rapidly transition between these binding modes using conventional sampling, then this can present a problem for free energy calculations, as the result obtained will depend on the binding mode of the ligand in the initial structure. Furthermore, this problem is compounded by the fact that the set of accessible binding modes may not be known *a priori.* There are therefore two problems that need to be solved: the determination of binding modes, and the inclusion of their impact in the free energy calculation.

The impact of the ligand binding mode was noted during a study on ACK1 by Granadino-Roldán *et al.* who showed that free energies calculated using binding modes obtained from docking do not correlate as favorably to experiment as the manually chosen ones.10 Similarly, a study by Cappel *et al.* on five different protein-ligand systems demonstrated that different methods of obtaining initial binding modes can in some cases result in strikingly different correlations against experiment, and they found that a docking procedure based on maximum common substructure (MCS) constraints exhibits most consistent performance on average.11

If the ligand of interest is known to exhibit multiple physically relevant binding modes, one way of taking all of them into account in evaluating binding affinity is to use each one as a starting point for a separate AFE calculation. These free energy values can afterwards be used to obtain a weighted average.12 However, the computational cost of this approach would quickly increase with the number of binding modes. A somewhat similar approach would be to calculate the free energy change of interest for one binding mode, and then to correct for all other binding modes using the free energy differences between the binding modes. If the free energy difference between the binding modes could be calculated more rapidly than the additional full free energy perturbation, this might be a more efficient route to a well-sampled result.

A method which offers solutions to each of the problems outlined in this section is the BLUES (Binding modes of Ligands Using Enhanced Sampling) method presented by Gill *et al.*9 This approach uses Monte Carlo sampling to propose changes in ligand binding mode during a simulation, making use of the nonequilibrium candidate Monte Carlo (NCMC) theory,13 which allows the environment to relax in response to the proposed change. The BLUES method allows the transitions between binding modes to be rapidly sampled, and therefore offers a convenient route to the identification of ligand binding modes and their relative populations.9,14–16The BLUES package has also been applied successfully to the sampling of protein sidechains17and water binding (discussed further in the following section).18

**Binding site hydration**

The displacement of buried water molecules, and the corresponding entropy increase, can be very beneficial in drug design,19 but can also present a significant practical issue for AFE calculations. The displacement/addition of a water site can be included as an extra stage before/after ligand perturbation, but, as demonstrated by Michel *et al.*, the result can be sensitive to the order of these steps, as the perturbation may be distorted by the incorrect presence or absence of a water site – one case was noted where a perturbation of around –6 kcal/mol in the absence of a key water can be around +14 kcal/mol if the water is present.20 There are two distinct, but closely related methods which can rigorously treat differences in water during free energy calculations: grand canonical Monte Carlo (GCMC)21–24 and the Monte Carlo translation method proposed by Ben-Shalom *et al*.25 The former involves inserting and deleting water molecules from a region of interest, via simulations at constant chemical potential, and the latter involves attempting explicit translations of waters between binding sites and bulk regions of the simulation. The latter approach has been implemented in the BLUES package, using NCMC to aid the acceptance of these moves.18 Both GCMC26–29 and the Monte Carlo translation method30 have been used to reduce the dependence of AFE calculations on the initial hydration state of the protein binding site, thereby improving reproducibility.

**Force fields**

**Protein/ligand force field**

Although the choice of protein force field can in principle impact the calculated binding free energy, the ligand force field is commonly considered to be the more important factor, owing to the larger number of atom types that need to be parametrized compared to the twenty standard amino acids. In addition, since the ligand is the part of the system which is perturbed in an AFE calculation, the resulting free energies are likely to be most sensitive to its parameters. Combined with the historical practice of using compatible protein/ligand force fields (e.g. ff14SB31/GAFF32 and CHARMM3633/CGenFF34,35), this means that the choice of protein force field is usually determined by the choice of ligand force field in practice. Moreover, while the effect of the ligand force field on free energy calculations can be separated from the protein force field in e.g. a solvation free energy calculation, this cannot be done in a binding free energy calculation, where it will be nonetheless assumed that the ligand force field is the main source of variability.

When comparing different force fields, it is rarely the case that all of them are implemented in a single MD engine. This inevitably introduces another layer of difficulty, as multiple MD engines need to be used to perform the study. In these cases it is difficult to decouple the effect of the force field parameters from other implementation differences between MD engines. Despite this difficulty, studies comparing different ligand force fields have been performed. For example, Vassetti *et al.* showed that solvation free energies calculated using either OPLS-AA or GAFF2 performed comparably against experiment on average but there still were significant differences between each of the calculated values, with a mean absolute difference of $\~$3.5 kcal/mol.36 Similarly, GAFF was also compared against OPLS3e, where the latter displayed better agreement with experiment and the correlation (R2) between results from the two force fields was system-dependent, ranging from $\~$0.6 to $\~$0.9.7

It transpires that sensitivity towards force field parameters also extends to minute differences in the parameters of the same force field. For example, Rocklin *et al.* demonstrated that small variations in the nonbonded parameters, such as charge differences of more than 0.02 e, can result in significant free energy changes of more than 1 kcal/mol. 37 These results are especially relevant to charge derivation methods, which are notorious for their dependence on the ligand conformation. In view of this, Manzoni and Ryde compared different charge derivation methods for ligands bound to galectin-3C using different starting ligand geometries and found that these can result in binding free energy discrepancies of more than 1 kcal/mol.38 Comparison to experiment was also inconsistent, with the RESP method39 generating the datasets with both highest and lowest correlation against experimental values.

Addressing the significant sensitivity of binding and solvation free energies to the force field functional form and its parameters is not trivial, owing to the inherent limitations of choosing a particular approximate functional form over another. Nevertheless, reproducibility between different force fields is expected to be higher if they are constantly updated to improve their performance against experiment in edge cases. This can be achieved by using bespoke force fields with a suitable level of quantum theory, such as QUBE40, or force fields which are constantly updated with extra parameters to handle edge cases, such as OPLS341 and OpenFF42. It is even more challenging to increase the reproducibility of the charge derivation methods, given the inherent limitations of atom-centered point charges. These can be partially circumvented by using force fields with a more sophisticated treatment of the electrostatic interactions, such as AMOEBA43, but these currently remain of limited utility to drug discovery due to their high computational cost.

**Water force field**

Additionally, the water model can also impact the results obtained. Although implicit/continuum solvent models44 can appear desirable, owing to the rapid convergence they offer, it has been demonstrated by Michel *et al.* that their usage can lead to inaccuracies in systems for which binding site waters are important.45 A GCMC study of 35 bromodomain proteins by Aldeghi *et al.* showed that whilst the trends in water binding free energies across the dataset were similar for the TIP3P, TIP4P and SPC water models46, the absolute values of the binding free energy can differ significantly.47 A related problem is the fact that water models are parametrized to reproduce bulk water properties,46 whereas protein binding sites may have significantly different electric fields48 to bulk water, causing the simulated water molecules to be under- or over-polarized. On this topic, Yin *et al.* developed the Bind3P water model by modifying the Lennard-Jones parameters of TIP3P, which was able to perform better than TIP3P for both host-guest binding free energies and hydration free energy calculations.49

**Hydrogen atom placement**

As hydrogen atoms are rarely assigned during the solution of a crystal structure model, this step must typically be carried out by modelers, prior to simulation. The majority of hydrogens in the system are non-labile (such as those attached to carbon atoms), and this is a simple task. However, this is more complex for residues with pKa values close to the intended pH (often taken as around 7.4), such as histidine for example, especially as pKa values can vary somewhat in a protein interior – it should be noted that this could also be relevant for other residues if the protein environment shifts the pKa sufficiently (such as the well-known Asp25 dyad in HIV-1 protease50). The decisions made regarding the protonation states of residues which are proximal to the perturbation of interest could have a significant impact on the resulting free energy, especially if the hydrogen bonding involving the perturbation is uncertain.

An alternative approach is the use of constant pH molecular dynamics simulations. Here, the various protonation states of protonatable residues are sampled during a simulation via the semi-grand canonical ensemble (in which the number of particles is constant, but particles can change identity). Here, changes in protonation state can be attempted using nonequilibrium Monte Carlo proposals, which are accepted based on both the work done and a reference pKa for the site of interest.51,52 Whilst constant pH simulations do not appear to have yet been used alongside alchemical free energy calculations, such an approach could be of use when the ligand or binding site residues can access multiple protonation states – especially if the relative probabilities of these protonation states might be expected to be dependent on the ligand perturbation. However, the additional cost incurred by including the protonation state sampling would have to be less than that of running additional free energy calculations starting from the alternative protonation states, and combining the resulting calculated free energies.

It is also possible to explicitly handle the protonation states and tautomers of interest – most notably, those of the ligand. A study by de Oliveira *et al.* combined the relative binding free energies of different tautomers and protonation states of ligands bound to kinesin spindle protein and factor Xa.53 To achieve this, they calculated the expected pKa values of the protonation site of interest using an *ab initio* approach. Even though their results showed consistent improvement against experiment, the computational cost of this approach increases exponentially with the number of titratable sites and is therefore only reserved for critical protonation sites in practice.

**Salt concentration**

Often in molecular simulations, researchers attempt to replicate the ionic strength of experimental solutions by including an appropriate concentration of salt in the simulation. There are several approaches to do this, one of which is to simply place the number of ion pairs in the simulation box which would correspond to the simulation volume at the desired concentration. However, this does not account for the fact that a large portion of this volume is occupied by the protein and therefore inaccessible to the ions, meaning that the effective concentration of the salt in the bulk solution is much higher than intended. Two alternatives to this are to determine the number of ions based on the solvent accessible volume (by subtracting the estimated volume excluded by the protein), or to calculate the ratio of ions to water molecules at the desired concentration. A problem that remains, however, is that the ion concentration, even if it correctly represents the experimental concentration, is fixed. The concentration reported in experimental measurements is an average over a macroscopic volume, so the concentrations observed instantaneously within a volume the size of a typical simulation would be expected to fluctuate.

A solution to each of these problems was proposed by Ross *et al*.,54 whose *saltswap* method employs simulations in the semi-grand canonical ensemble, in which pairs of water molecules are exchanged with ion pairs. This allows the number of ion pairs within the simulation to fluctuate according to the equilibrium distribution at a particular macroscopic concentration, thereby offering a solution to both of the problems described: the number of ions in the simulations is not fixed and need not be determined *a priori*. This approach might offer more reproducible results, as it simply requires a definition of the macroscopic salt concentration. They found that the mean concentration observed during these rigorous simulations was not well-reproduced by the methods involving a fixed number of ions.54

The choice of force field parameters for the ions present in the simulation might also impact the results obtained from free energy calculations. This is especially relevant if there are any biologically relevant ions present in the binding site of a protein. It has also been pointed out by Leontyev and Stuchebrukhov that the lack of polarization of molecular simulations can result in inadequate electrostatic screening of ions by the solvent. For aqueous simulations, this can be accounted for by scaling the ion charges by a factor of ~0.7.55

**Simulation Details**

**Sampling time**

As mentioned above, many of the initial choices during system preparation (particularly those of initial coordinates), should not, in principle, affect the true ensemble average. However, biologically relevant timescales (milliseconds to seconds) are beyond the reach of most modern computing capabilities. Moreover, alchemical free energy calculations often need multiple simulations to obtain a converged free energy estimate, meaning that one can only dedicate a fraction of the allocated computer time to a single λ window. Consequently, the length of each λ window is often chosen in practice to be in the range of 1 – 5 ns, especially in commercial applications56,57.

While it is obvious that longer simulation times provide the researcher with more highly decorrelated structures and access to molecular motions that are inaccessible at shorter timescales, drug discovery applications benefit most from high throughput, since even short alchemical calculations are expensive. Therefore, computational free energy studies have historically focused on direct comparison to experiment rather than measuring the short-timescale bias with respect to the true ensemble average predicted by the force field.

A large-scale study by Fratev and Sirimulla investigated the quality of the free energy values with respect to equilibration time and the simulation time.58 A key point in their paper is that there is a practical tradeoff in terms of both equilibration and simulation time, with short simulations comparing unfavorably to experiment and long simulations resulting in low throughput. They found that an optimal procedure involves a pre-REST equilibration protocol of two independent 10 ns runs followed by 8 ns sampling time per λ window. This protocol results in an approximately two-fold decrease in MAE with respect to experiment across all five protein systems studied. Their results suggest that while apparent convergence in the sampling stage is not difficult to obtain, prolonged equilibration is crucial for exploring crucial slow modes of motion. Therefore, proteins with higher levels of structural flexibility benefit more from these extended protocols.

Nevertheless, other studies have found that extended protocols do not necessarily result in better agreement with experiment. For instance, Wan *et al.* showed that a tenfold increase in sampling from 4 to 40 ns can significantly reduce correlation with experiment, while any improvement is on average negligible regardless of the sampling method used.59 In many cases the authors observed a significant shift in predicted free energies after extending the length of the simulation, often reaching 1 kcal/mol, indicating improved sampling offset by an insufficiently accurate force field model. Similar observations have been made by Suruzhon *et al.*, who found that the sampling differences stemming from a 20 ns equilibration can, in some cases, reach 5 kcal/mol.6 This notwithstanding, the average agreement with experiment was not significantly improved, once again suggesting that the benefit of the increased sampling is only observed when the force field model is adequate for the system.

**Free energy estimator**

The choice of the free energy estimator can also have an impact on the obtained free energy value. The most widely used equilibrium free energy estimators are the Zwanzig equation60 (FEP), the Bennett acceptance ratio61 (BAR) and its multistate generalization62 (MBAR), and thermodynamic integration (TI). For a perturbation involving $N$ total λ windows, FEP requires $N-1$ simulations, unlike all other methods, which require the full set of $N$ simulations. TI, on the other hand, needs only $N$ energy evaluations per unit time across all λ windows, compared to $2N-2$ for FEP, $3N-2$ for BAR and $N^{2}$ for MBAR. Consequently, even though MBAR has been proven to be the asymptotically statistically optimal estimator for$ N$ λ windows (reducing to BAR at $N=2$)62, BAR and TI are still commonly used in the literature because of their lower computational requirements, while simultaneously providing sufficiently good accuracy in many cases.

Although it is usually assumed that the above estimators will eventually converge to the same value with infinite sampling, this is not the case for TI, which also requires infinitely many intermediate λ windows for asymptotic convergence. While these conceptual differences are not necessarily practically significant, discrepancies between the estimated free energy values do arise in some cases. For example, a study by de Ruiter *et al.* has demonstrated that there can be strikingly large differences between BAR and TI estimates in protein-ligand binding free energy calculations, in some cases reaching 3 kcal/mol.63 Moreover, TI is also dependent on the integration procedure used, which can result in differences of 1 kcal/mol.63 Nevertheless, the authors observed that increasing the number of λ windows to 21 makes most of these discrepancies negligible, showing that while TI is more sensitive to the shape of the free energy profile, it is still systematically improvable in practice.

These observations are in accordance with an earlier publication by Shirts and Pande, which also showed that BAR is expected to significantly outperform TI and FEP in most practical use cases.64 However, more recent developments have shown that TI can perform sufficiently well in practice with a carefully designed protocol65. Nevertheless, BAR and MBAR remain the free energy estimators that do not require any additional user input (c.f. choosing FEP direction or TI integration method), making them the most reproducible free energy methods.

**Independent repeats**

Increasing the number of independent simulation repeats reduces the standard error of the free energy estimate, thereby increasing its statistical certainty and hence reproducibility. However, the free energy variance can often be underestimated due to undersampling, meaning that its associated standard error can appear to increase with a higher number of replicates. This is, of course, only an apparent effect and the true standard error of the free energy estimate will always decrease with a large number of samples. This was demonstrated in a study carried out by Bhati *et al.*, who reported that ensemble simulations improve the reliability of the estimated free energy uncertainty.66 This approach involves running a number (5-10) of independent repeats, and then combining all samples before estimating the free energy,67 as opposed to carrying out fewer, longer repeats for an equivalent computational cost.

However, there is always a practical tradeoff between the number of repeats and the resources allotted to a single binding free energy calculation6. While it is important to obtain an estimate of the free energy variance, it is also highly desirable to obtain effective decorrelation from the initial coordinates and sample binding rare events, which can only be achieved with enhanced sampling and/or longer timescales. It is therefore not clear where the optimal balance lies between longer simulations and more repeats – the definition of “optimal” is also somewhat ambiguous in this scenario. In any case, several replicate simulations should be performed in practice to measure the repeatability of the results.

**Softcore potential**

When performing alchemical free energy calculations, the choice of the functional form of the energy coupling between the end states is arbitrary. However, it is desirable to choose a functional form which minimizes the free energy variance over λ space. In practice, this is commonly achieved using softcore potentials, which soften the potential energy singularities of the alchemical atoms. Softcore potentials are most commonly used with van der Waals interactions, but can also be used with electrostatic interactions. The choice of the softcore potential, its parameters, and the protocol of perturbing the bonded, van der Waals, and electrostatic interactions can all affect the free energy estimate in nonobvious ways.

Arguably the most widely used softcore potential uses an effective radius $r\_{ij,eff}$ between two atoms $i$ and $j$, which is related to the real radius $r\_{ij}$ in the following way:

$$r\_{ij,eff}=\left(σ\_{ij}^{c}αλ^{b}+r\_{ij}^{c}\right)^{\frac{1}{c}}$$

Here $α$ is a continuous parameter, $b$ and $c$ are discrete parameters, and $σ\_{ij}$ is a force field parameter (the average particle “size”)68. The nature of the optimal parameters has also been investigated. For example, Steinbrecher *et al.* identified an acceptable range of softcore potential values, while demonstrating that the exact value does not significantly affect the free energy estimate itself, only its variance.69 However, de Ruiter *et al.* observed significant differences between some softcore parameter combinations, in some cases reaching 1 kcal/mol discrepancies.63 Nevertheless, these are also dependent on the number of intermediate λ windows, meaning that with an insufficient amount of intermediate states, discrepancies of 2 kcal/mol can be observed using BAR and > 10 kcal/mol using TI.

The viability of simultaneously performing the van der Waals and electrostatic perturbations has also been investigated in several studies. For instance, Steinbrecher *et al*. showed that both split and unified protocols result in consistent free energy estimates.69 Although Loeffler *et al.* obtained results which were largely consistent with this notion, they observed in some cases free energy discrepancies of up to 1.5 kcal/mol.70 Even though these differences appeared well-converged, the authors noted that the split protocol is often more desirable than the unified, since electrostatic softcore potentials often introduce irregularities in the free energy profile, meaning that the free energies are more easily reproducible.

**Other**

Alchemical free energy calculations are often modelled on the basis of either a canonical (NVT), or an isothermal-isobaric (NPT) ensemble. As such, the choice and the parameters of the corresponding integrators, thermostats and barostats can also potentially affect the resulting free energy estimates. For example, a deficiency of the Berendsen barostat was found during the course of the SAMPL6 challenge, resulting in non-negligible sampling artifacts.71 More generally, it is well-known that velocity- and pressure-rescaling algorithms do not correctly sample from the corresponding thermodynamic distributions72 and they should be avoided in MD simulations. In addition, different integrator splittings can give rise to significantly different sampling distributions73. Although these can be corrected by adding a Metropolization step74 which ensures stationarity of the Boltzmann distribution, many practitioners still use rescaling algorithms or integrators without Metropolization, thereby hindering reproducibility.

Finally, the MD engine of choice can also have a non-negligible impact on the calculated free energies. Quantifying the extent of this impact is nontrivial, however, not least because different MD engines implement different thermostats, barostats, integrators, softcore potentials, etc. For instance, a study by Loeffler *et al.*70 found a reproducibility limit of 0.2 kcal/mol between relative free energies of solvation and similar discrepancies of up to 1 kcal/mol were further observed in a host-guest system during the SAMPL6 challenge.71 Although narrowing down the reason for these inconsistencies remains largely speculative, they are at least partially explained by code issues, some of which have notably been fixed since75. It is therefore important to keep oneself up to date with major bug fixes, as well as test any freshly installed code on basic benchmarks. On the developer side, this also means that any major bug fixes should come with additional unit and integration tests to prevent the accidental reintroduction of the bug.

**Summary**

In this chapter, we have outlined most of the practical choices relevant to AFE calculations. A general trend in the various discussions is that many choices made by the researcher can significantly affect the calculated free energy values and it is therefore important to carefully consider as many of these choices as possible when designing a molecular modelling project in drug discovery.

For example, using a combination of enhanced sampling methods to calculate binding free energies in e.g. triplicates will convert many of the hidden structural biases into observable statistical variance which will provide the researcher with realistic confidence intervals of the results. Combining enhanced sampling methods is also becoming an increasingly easier task with software projects such as OpenMM76 and PLUMED,77 which help the researcher customize their own study using a specific combination of enhanced sampling algorithms with relatively little effort.

Of significant interest are the decisions which cannot be immediately resolved by longer simulations. Some of them can still be mapped onto an ensemble of Hamiltonians (e.g. protonation states), while further studies and developments are needed to reduce discrepancies in others (e.g. choice of software and force field).

Another simple route to improving reproducibility would be for the community to impose stricter standards as to how free energy protocols and analyses are reported, such that an interested reader would be able to replicate the results from a publication. A similar, but alternative, approach could be to normalize the inclusion of input files and analysis scripts in the supporting information, which would facilitate reproduction of protocols by others. More widespread imposition of requirements such as these by journals and/or reviewers would be of significant benefit to their readers.

Finally, we should like to emphasize the importance of large-scale molecular modelling challenges, such as SAMPL.71 These can unambiguously identify robustness problems based on a large set of molecular simulations. These problems can then be investigated and addressed in further depth during subsequent studies.

**References**

(1) Singh, N.; Warshel, A. Absolute Binding Free Energy Calculations: On the Accuracy of Computational Scoring of Protein-Ligand Interactions. *Proteins Struct. Funct. Bioinforma.* **2010**, NA-NA. https://doi.org/10.1002/prot.22687.

(2) Hahn, D. F.; Bayly, C. I.; Macdonald, H. E. B.; Chodera, J. D.; Mey, A. S. J. S.; Mobley, D. L.; Benito, L. P.; Schindler, C. E. M.; Tresadern, G.; Warren, G. L. *Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks*; preprint; 2021.

(3) Cavasotto, C. N.; Phatak, S. S. Homology Modeling in Drug Discovery: Current Trends and Applications. *Drug Discov. Today* **2009**, *14* (13), 676–683. https://doi.org/10.1016/j.drudis.2009.04.006.

(4) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46* (W1), W296–W303. https://doi.org/10.1093/nar/gky427.

(5) Haddad, Y.; Adam, V.; Heger, Z. Ten Quick Tips for Homology Modeling of High-Resolution Protein 3D Structures. *PLOS Comput. Biol.* **2020**, *16* (4), e1007449. https://doi.org/10.1371/journal.pcbi.1007449.

(6) Suruzhon, M.; Bodnarchuk, M. S.; Ciancetta, A.; Viner, R.; Wall, I. D.; Essex, J. W. Sensitivity of Binding Free Energy Calculations to Initial Protein Crystal Structure. *J. Chem. Theory Comput.* **2021**, *17* (3), 1806–1821. https://doi.org/10.1021/acs.jctc.0c00972.

(7) Pérez-Benito, L.; Casajuana-Martin, N.; Jiménez-Rosés, M.; van Vlijmen, H.; Tresadern, G. Predicting Activity Cliffs with Free-Energy Perturbation. *J. Chem. Theory Comput.* **2019**, *15* (3), 1884–1895. https://doi.org/10.1021/acs.jctc.8b01290.

(8) Lim, N. M.; Wang, L.; Abel, R.; Mobley, D. L. Sensitivity in Binding Free Energies Due to Protein Reorganization. *J. Chem. Theory Comput.* **2016**, *12* (9), 4620–4631. https://doi.org/10.1021/acs.jctc.6b00532.

(9) Gill, S. C.; Lim, N. M.; Grinaway, P. B.; Rustenburg, A. S.; Fass, J.; Ross, G. A.; Chodera, J. D.; Mobley, D. L. Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *J. Phys. Chem. B* **2018**, *122* (21), 5579–5598. https://doi.org/10.1021/acs.jpcb.7b11820.

(10) Granadino-Roldán, J. M.; Mey, A. S. J. S.; Pérez González, J. J.; Bosisio, S.; Rubio-Martinez, J.; Michel, J. Effect of Set up Protocols on the Accuracy of Alchemical Free Energy Calculation over a Set of ACK1 Inhibitors. *PloS One* **2019**, *14* (3), e0213217–e0213217. https://doi.org/10.1371/journal.pone.0213217.

(11) Cappel, D.; Jerome, S.; Hessler, G.; Matter, H. Impact of Different Automated Binding Pose Generation Approaches on Relative Binding Free Energy Simulations. *J. Chem. Inf. Model.* **2020**, *60* (3), 1432–1444. https://doi.org/10.1021/acs.jcim.9b01118.

(12) Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical Free Energy Calculations for Drug Discovery. *J. Chem. Phys.* **2012**, *137* (23), 230901. https://doi.org/10.1063/1.4769292.

(13) Nilmeier, J. P.; Crooks, G. E.; Minh, D. D. L.; Chodera, J. D. Nonequilibrium Candidate Monte Carlo Is an Efficient Tool for Equilibrium Simulation. *Proc. Natl. Acad. Sci.* **2011**, *108* (45), E1009–E1018. https://doi.org/10.1073/pnas.1106094108.

(14) Jandova, Z.; Gill, S. C.; Lim, N. M.; Mobley, D. L.; Oostenbrink, C. Binding Modes and Metabolism of Caffeine. *Chem. Res. Toxicol.* **2019**, *32* (7), 1374–1383. https://doi.org/10.1021/acs.chemrestox.9b00030.

(15) Lim, N. M.; Osato, M.; Warren, G. L.; Mobley, D. L. Fragment Pose Prediction Using Non-Equilibrium Candidate Monte Carlo and Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2020**, *16* (4), 2778–2974. https://doi.org/10.1021/acs.jctc.9b01096.

(16) Sasmal, S.; Gill, S. C.; Lim, N. M.; Mobley, D. L. Sampling Conformational Changes of Bound Ligands Using Nonequilibrium Candidate Monte Carlo and Molecular Dynamics. *J. Chem. Theory Comput.* **2020**, *16* (3), 1854–1865. https://doi.org/10.1021/acs.jctc.9b01066.

(17) Burley, K. H.; Gill, S. C.; Lim, N. M.; Mobley, D. L. Enhancing Side Chain Rotamer Sampling Using Nonequilibrium Candidate Monte Carlo. *J. Chem. Theory Comput.* **2019**, *15* (3), 1848–1862. https://doi.org/10.1021/acs.jctc.8b01018.

(18) Bergazin, T. D.; Ben-Shalom, I. Y.; Lim, N. M.; Gill, S. C.; Gilson, M. K.; Mobley, D. L. Enhancing Water Sampling of Buried Binding Sites Using Nonequilibrium Candidate Monte Carlo. *J. Comput. Aided Mol. Des.* **2021**, *35* (2), 167–177. https://doi.org/10.1007/s10822-020-00344-8.

(19) Ladbury, J. E. Just Add Water! The Effect of Water on the Specificity of Protein-Ligand Binding Sites and Its Potential Application to Drug Design. *Chem. Biol.* **1996**, *3* (12), 973–980. https://doi.org/10.1016/S1074-5521(96)90164-7.

(20) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *J. Am. Chem. Soc.* **2009**, *131* (42), 15403–15411. https://doi.org/10.1021/ja906058w.

(21) Woo, H.-J.; Dinner, A. R.; Roux, B. Grand Canonical Monte Carlo Simulations of Water in Protein Environments. *J. Chem. Phys.* **2004**, *121* (13), 6392–6400. https://doi.org/10.1063/1.1784436.

(22) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. Water Sites, Networks, And Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.* **2015**, *137* (47), 14930–14943. https://doi.org/10.1021/jacs.5b07940.

(23) Ross, G. A.; Bruce Macdonald, H. E.; Cave-Ayland, C.; Cabedo Martinez, A. I.; Essex, J. W. Replica-Exchange and Standard State Binding Free Energies with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2017**, *13* (12), 6373–6381. https://doi.org/10.1021/acs.jctc.7b00738.

(24) Samways, M. L.; Bruce Macdonald, H. E.; Essex, J. W. Grand: A Python Module for Grand Canonical Water Sampling in OpenMM. *J. Chem. Inf. Model.* **2020**, *60* (10), 4436–4441. https://doi.org/10.1021/acs.jcim.0c00648.

(25) Ben-Shalom, I. Y.; Lin, C.; Kurtzman, T.; Walker, R. C.; Gilson, M. K. Simulating Water Exchange to Buried Binding Sites. *J. Chem. Theory Comput.* **2019**, *15* (4), 2684–2691. https://doi.org/10.1021/acs.jctc.8b01284.

(26) Deng, Y.; Roux, B. Computation of Binding Free Energy with Molecular Dynamics and Grand Canonical Monte Carlo Simulations. *J. Chem. Phys.* **2008**, *128* (11), 115103. https://doi.org/10.1063/1.2842080.

(27) Bruce Macdonald, H. E.; Cave-Ayland, C.; Ross, G. A.; Essex, J. W. Ligand Binding Free Energies with Adaptive Water Networks: Two-Dimensional Grand Canonical Alchemical Perturbations. *J. Chem. Theory Comput.* **2018**, *14* (12), 6586–6597. https://doi.org/10.1021/acs.jctc.8b00614.

(28) Ross, G. A.; Russell, E.; Deng, Y.; Lu, C.; Harder, E. D.; Abel, R.; Wang, L. Enhancing Water Sampling in Free Energy Calculations with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2020**, *16* (10), 6061–6076. https://doi.org/10.1021/acs.jctc.0c00660.

(29) Wahl, J.; Smieško, M. Assessing the Predictive Power of Relative Binding Free Energy Calculations for Test Cases Involving Displacement of Binding Site Water Molecules. *J. Chem. Inf. Model.* **2019**, *59* (2), 754–765. https://doi.org/10.1021/acs.jcim.8b00826.

(30) Ben-Shalom, I. Y.; Lin, Z.; Radak, B. K.; Lin, C.; Sherman, W.; Gilson, M. K. Accounting for the Central Role of Interfacial Water in Protein–Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2020**, *16* (12), 7883–7894. https://doi.org/10.1021/acs.jctc.0c00785.

(31) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10/f7nfc3.

(32) 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* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(33) Huang, J.; MacKerell Jr, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* **2013**, *34* (25), 2135–2145. https://doi.org/10.1002/jcc.23354.

(34) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52* (12), 3144–3154. https://doi.org/10.1021/ci300363c.

(35) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52* (12), 3155–3168. https://doi.org/10.1021/ci3003649.

(36) Vassetti, D.; Pagliai, M.; Procacci, P. Assessment of GAFF2 and OPLS-AA General Force Fields in Combination with the Water Models TIP3P, SPCE, and OPC3 for the Solvation Free Energy of Druglike Organic Molecules. *J. Chem. Theory Comput.* **2019**, *15* (3), 1983–1995. https://doi.org/10.1021/acs.jctc.8b01039.

(37) Rocklin, G. J.; Mobley, D. L.; Dill, K. A. Calculating the Sensitivity and Robustness of Binding Free Energy Calculations to Force Field Parameters. *J. Chem. Theory Comput.* **2013**, *9* (7), 3072–3083. https://doi.org/10.1021/ct400315q.

(38) Manzoni, F.; Ryde, U. Assessing the Stability of Free-Energy Perturbation Calculations by Performing Variations in the Method. *J. Comput. Aided Mol. Des.* **2018**, *32* (4), 529–536. https://doi.org/10.1007/s10822-018-0110-5.

(39) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269–10280. https://doi.org/10.1021/j100142a004.

(40) Allen, A. E. A.; Robertson, M. J.; Payne, M. C.; Cole, D. J. Development and Validation of the Quantum Mechanical Bespoke Protein Force Field. *ACS Omega* **2019**, *4* (11), 14537–14550. https://doi.org/10.1021/acsomega.9b01769.

(41) 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* (1), 281–296. https://doi.org/10.1021/acs.jctc.5b00864.

(42) Qiu, Y.; Smith, D.; Boothroyd, S.; Jang, H.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C.; Rizzi, A.; Lucas, X.; Tjanaka, B.; Shirts, M. R.; Gilson, M.; Chodera, J.; Bayly, C. I.; Mobley, D.; Wang, L.-P. *Development and Benchmarking of Open Force Field v1.0.0, the Parsley Small Molecule Force Field*; preprint; 2020. https://doi.org/10.26434/chemrxiv.13082561.v1.

(43) Ren, P.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107* (24), 5933–5947. https://doi.org/10.1021/jp027815+.

(44) Massova, I.; Kollman, P. A. Combined Molecular Mechanical and Continuum Solvent Approach (MM-PBSA/GBSA) to Predict Ligand Binding. *Perspect. Drug Discov. Des.* **2000**, *18* (1), 113–135. https://doi.org/10.1023/A:1008763014207.

(45) Michel, J.; Verdonk, M. L.; Essex, J. W. Protein-Ligand Binding Affinity Predictions by Implicit Solvent Simulations: A Tool for Lead Optimization? *J. Med. Chem.* **2006**, *49* (25), 7427–7439. https://doi.org/10.1021/jm061021s.

(46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10/dg9sq8.

(47) Aldeghi, M.; Ross, G. A.; Bodkin, M. J.; Essex, J. W.; Knapp, S.; Biggin, P. C. Large-Scale Analysis of Water Stability in Bromodomain Binding Pockets with Grand Canonical Monte Carlo. *Commun. Chem.* **2018**, *1* (1), 19. https://doi.org/10.1038/s42004-018-0019-x.

(48) Bradshaw, R. T.; Dziedzic, J.; Skylaris, C.-K.; Essex, J. W. The Role of Electrostatics in Enzymes: Do Biomolecular Force Fields Reflect Protein Electric Fields? *J. Chem. Inf. Model.* **2020**, *60* (6), 3131–3144. https://doi.org/10.1021/acs.jcim.0c00217.

(49) Yin, J.; Henriksen, N. M.; Muddana, H. S.; Gilson, M. K. Bind3P: Optimization of a Water Model Based on Host–Guest Binding Data. *J. Chem. Theory Comput.* **2018**, *14* (7), 3621–3632. https://doi.org/10.1021/acs.jctc.8b00318.

(50) Trylska, J.; Antosiewicz, J.; Geller, M.; Hodge, C. N.; Klabe, R. M.; Head, M. S.; Gilson, M. K. Thermodynamic Linkage between the Binding of Protons and Inhibitors to HIV-1 Protease. *Protein Sci.* **1999**, *8* (1), 180–195. https://doi.org/10.1110/ps.8.1.180.

(51) Chen, Y.; Roux, B. Constant-PH Hybrid Nonequilibrium Molecular Dynamics–Monte Carlo Simulation Method. *J. Chem. Theory Comput.* **2015**, *11* (8), 3919–3931. https://doi.org/10.1021/acs.jctc.5b00261.

(52) Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B. Constant-PH Molecular Dynamics Simulations for Large Biomolecular Systems. *J. Chem. Theory Comput.* **2017**, *13* (12), 5933–5944. https://doi.org/10.1021/acs.jctc.7b00875.

(53) de Oliveira, C.; Yu, H. S.; Chen, W.; Abel, R.; Wang, L. Rigorous Free Energy Perturbation Approach to Estimating Relative Binding Affinities between Ligands with Multiple Protonation and Tautomeric States. *J. Chem. Theory Comput.* **2019**, *15* (1), 424–435. https://doi.org/10.1021/acs.jctc.8b00826.

(54) Ross, G. A.; Rustenburg, A. S.; Grinaway, P. B.; Fass, J.; Chodera, J. D. Biomolecular Simulations under Realistic Macroscopic Salt Conditions. *J. Phys. Chem. B* **2018**, *122* (21), 5466–5486. https://doi.org/10.1021/acs.jpcb.7b11734.

(55) Leontyev, I.; Stuchebrukhov, A. Accounting for Electronic Polarization in Non-Polarizable Force Fields. *Phys. Chem. Chem. Phys.* **2011**, *13* (7), 2613. https://doi.org/10.1039/c0cp01971b.

(56) 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.

(57) 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.

(58) Fratev, F.; Sirimulla, S. An Improved Free Energy Perturbation FEP+ Sampling Protocol for Flexible Ligand-Binding Domains. *Sci. Rep.* **2019**, *9* (1), 16829. https://doi.org/10.1038/s41598-019-53133-1.

(59) Wan, S.; Tresadern, G.; Pérez‐Benito, L.; Vlijmen, H.; Coveney, P. V. Accuracy and Precision of Alchemical Relative Free Energy Predictions with and without Replica‐Exchange. *Adv. Theory Simul.* **2020**, *3* (1), 1900195. https://doi.org/10.1002/adts.201900195.

(60) Zwanzig, R. W. High‐Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426. https://doi.org/10.1063/1.1740409.

(61) 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.

(62) 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.

(63) de Ruiter, A.; Boresch, S.; Oostenbrink, C. Comparison of Thermodynamic Integration and Bennett Acceptance Ratio for Calculating Relative Protein-Ligand Binding Free Energies. *J. Comput. Chem.* **2013**, *34* (12), 1024–1034. https://doi.org/10.1002/jcc.23229.

(64) 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* (14), 144107. https://doi.org/10.1063/1.1873592.

(65) 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.

(66) 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* (6), 2867–2880. https://doi.org/10.1021/acs.jctc.7b01143.

(67) Coveney, P. V.; Wan, S. On the Calculation of Equilibrium Thermodynamic Properties from Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2016**, *18* (44), 30236–30240. https://doi.org/10.1039/C6CP02349E.

(68) 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* (3), 034114. https://doi.org/10.1063/1.3607597.

(69) Steinbrecher, T.; Joung, I.; Case, D. A. Soft-Core Potentials in Thermodynamic Integration: Comparing One- and Two-Step Transformations. *J. Comput. Chem.* **2011**, *32* (15), 3253–3263. https://doi.org/10.1002/jcc.21909.

(70) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14* (11), 5567–5582. https://doi.org/10.1021/acs.jctc.8b00544.

(71) 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* (5), 601–633. https://doi.org/10.1007/s10822-020-00290-5.

(72) Braun, E.; Moosavi, S. M.; Smit, B. Anomalous Effects of Velocity Rescaling Algorithms: The Flying Ice Cube Effect Revisited. *J. Chem. Theory Comput.* **2018**, *14* (10), 5262–5272. https://doi.org/10.1021/acs.jctc.8b00446.

(73) Fass, J.; Sivak, D. A.; Crooks, G. E.; Beauchamp, K. A.; Leimkuhler, B.; Chodera, J. D. Quantifying Configuration-Sampling Error in Langevin Simulations of Complex Molecular Systems. *Entropy* **2018**, *20* (5), 318. https://doi.org/10.3390/e20050318.

(74) Wagoner, J. A.; Pande, V. S. Reducing the Effect of Metropolization on Mixing Times in Molecular Dynamics Simulations. *J. Chem. Phys.* **2012**, *137* (21), 214105. https://doi.org/10.1063/1.4769301.

(75) Tsai, H.-C.; Tao, Y.; Lee, T.-S.; Merz, K. M.; York, D. M. Validation of Free Energy Methods in AMBER. *J. Chem. Inf. Model.* **2020**, *60* (11), 5296–5300. https://doi.org/10.1021/acs.jcim.0c00285.

(76) 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* (7), e1005659. https://doi.org/10/gbppkv.

(77) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613. https://doi.org/10.1016/j.cpc.2013.09.018.