Perturbation free-energy toolkit: automated alchemical topology builder and optimized simulation update scheme

Drazen Petrov[1]*

**corresponding author*

[1]Department of Material Sciences and Process Engineering, Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences Vienna, Muthgasse 18, A-1190 Vienna, Austria

KEYWORDS:

Molecular dynamics simulations, free energy calculations, alchemical perturbation, python package

ABSTRACT

Free-energy calculations play an important role in the application of computational chemistry to a range of fields, including protein biochemistry, rational drug design or material science. Importantly, the free energy difference is directly related to experimentally measurable quantities such as partition and adsorption coefficients, water activity and binding affinities. Among several techniques aimed at predicting the free-energy differences, perturbation approaches, involving alchemical transformation of one molecule into another through intermediate states, stand out as rigorous methods based on statistical mechanics. However, despite the importance of efficient and accurate free energy predictions, applicability of the perturbation approaches is still largely impeded by a number of challenges. This study aims at addressing two of them: 1) the definition of the perturbation path, i.e., alchemical changes leading to the transformation of one molecule to the other, and 2) determining the amount of sampling along the path to reach desired convergence. In particular, an automatic perturbation builder based on a graph matching algorithm is developed, that is able to identify the maximum common substructure of two molecules and provide the perturbation topologies suitable for free-energy calculations using GROMOS and GROMACS simulation packages. Moreover, it was used to calculate the changes in free energy of a set of post-translational modifications and analyze their convergence behavior. Different methods were tested, which showed that MBAR and extended thermodynamic integration (TI) in combination with MBAR show better performance as compared to BAR, extended TI with linear interpolation and plain TI. Also, a number of error estimators were explored and how they relate to the true error, estimated as the difference in free energy from an extensive set of simulation data. This analysis shows that most of the estimators provide only a qualitative agreement to the true error, with little quantitative predictive power.

This notwithstanding, the preformed analyses provided insight into the convergence of free-energy calculations, which allowed for development of an iterative update scheme for perturbation simulations that aims at minimizing the simulation time to reach the convergence, i.e., optimizing the efficiency. Importantly, this toolkit is made available online as an open-source python package (https://github.com/drazen-petrov/SMArt).

INTRODUCTION

Calculation of free-energy differences is one of the main objectives in computational chemistry as such differences characterize chemical processes, directly determining properties such as ligand binding affinities or partition coefficients. Perturbation free-energy calculations, involving alchemical transformation of one chemical into another via a pathway of unphysical intermediate states, present a rigorous approach derived from statistical mechanics.[1–12] Several such methods have been developed over the years, including for instance thermodynamic integration,[13] its extended version[14] or Bennett's acceptance ratio.[15] More recently, non-equilibrium techniques like the Crooks Gaussian intersection method[16] and the Jarzynski equality,[17] have also been applied. While more tractable than the direct simulations of the actual physical process (e.g. ligand binding), perturbation simulations are still computationally demanding, presenting one of the major impediments of their wider application.

The efficiency of different perturbation methods in various contexts have been studied.[16,18–21] In addition, the effects of the choice of intermediate states and exact coupling of the transformation to the Hamiltonian of the system through a coupling parameter λ have been explored.[22–26] Related to this, the transformation pathway depends on the definition of alchemical changes, which in turn might strongly affect the performance of the calculations. In particular, the dual topology approach replaces all atoms of one compound with the atoms of the other. Alternatively, only a subset of non-matching atoms can be perturbed, which is especially beneficial when compounds in question share the same scaffold. Performing free energy calculations using such an approach usually involves a cumbersome and often manual procedure of defining the perturbations, choosing intermediate states and the amount of sampling for simulations, followed by analysis of the collected data. On the other hand, several available tools

allow for automatization of some of the steps involved in the process, including the generation of perturbation topologies and simulation setup.[27–36]

In this study, two challenges related to the efficiency of perturbation free-energy calculations were addressed. In particular, an automated perturbation topology builder based on a graph-matching algorithm was developed allowing the user to find the maximum common substructure of two or a set of multiple compounds and define the perturbation accordingly. Secondly, this tool was used to generate perturbation topologies and to calculate free energies of four model post-translational modifications in the context of the histone H3 tail peptide and one of its readers Sp100C PHD finger domain.[37–40] In the process, the convergence behavior of these calculations was analyzed. Based on this analysis, an iterative update scheme of simulations along the perturbation path was proposed, attempting to automatically obtain converged free-energy differences for a minimum amount of simulation time. Finally, the toolkit is made available as an open source python package via a github repository (https://github.com/drazen-petrov/SMArt).

METHODS

Perturbation topology builder

The perturbation topology builder creates a definition of the perturbation pathway for a set of input molecular topologies (at least two), needed for free-energy calculations based on the maximal common substructure. The maximal common substructure (MCS) search for the two molecules involved in the perturbation is based on the VF algorithm for graph isomorphism matching.[41] It involves an iterative procedure in which in each step a pair of atoms, each belonging one of the compounds, is added to the current common substructure (current solution).

5

Upon adding a pair of atoms, the common part of molecular topologies is checked for non-matching forcefield parameters, which contribute to a score, based on user-defined penalty. At this point, a list of available pairs of atoms to be added in the current solution is updated, based on the first neighbors of the atoms in the current solution. A crucial part of this update is an estimate of the minimal penalization score that this current solution can achieve, according to which the list of available pairs of atoms is sorted. This ensures that solutions with low penalty scores are found early in the enumeration. When a current solution's minimal possible score is higher than a score of an already enumerated solution, this branch of enumeration is pruned. An initial point in the algorithm is a list of all available pairs of atoms, equaling n x m, where n and m stand for the number of atoms in each of the compounds. Importantly, this algorithm can also be simultaneously applied on a set of multiple topologies, where the resulting match represents the minimum structure of which each individual compound is a substructure, or simply put a common scaffold. This can be used to perform Enveloping Distribution Sampling (EDS)[42–44] or generate closed thermodynamic cycles on a set of multiple compounds.

The algorithm is implemented in python programming language and supports GROMOS and GROMACS file formats.

Perturbation simulations

Molecular dynamics simulations were performed using the GROMOS11[45] and GROMACS[46] simulation packages. The united-atom GROMOS force field, parameter set 54A8,[47–49] SPC explicit water[50] and 2 fs integration step were used. The temperature and the pressure were kept constant at 300 K and 1 bar using weak coupling with a relaxation time of 0.1 ps and 0.5 ps, respectively.[51] Pressure scaling was applied isotropically, with an isothermal compressibility of 4.575×10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$. A reaction-field contribution was added to the electrostatic

interactions and forces to account for a homogeneous medium with a dielectric permittivity of 61 outside the cutoff sphere. In simulations using the GROMOS11 molecular simulation package, a molecular pair-list was generated using a triple-range cutoff,[52] where nonbonded interactions up to a short range of 0.8 nm were calculated at every time step from a pair-list that was updated every 5 steps. Interactions up to a long-range cutoff of 1.4 nm were calculated at pair-list updates and kept constant in between. The SHAKE algorithm was used to constrain the bond lengths to their optimal values with a relative geometric accuracy of 10^{-4}.[53] In simulations performed using GROMACS simulation package, Verlet pair-list algorithm[54] was used, together with the LINCS algorithm[55] for constraining the bond lengths to their optimal values.

The above developed tool was used to define alchemical perturbations from native residues to their post-translationally modified forms (Table 1). A soft-core potential was used for perturbations of nonbonded interactions.[56] Free energies of these transformations of a small pentapeptide (GGXGG, where X stands for the affected residue with charge-neutral terminal caps) in the free state, i.e., in water, were performed with the GROMOS. The lysine trimethylation of a histone H3 tail peptide bound to the Sp100C PHD finger domain,[40] was simulated with the GROMACS simulation package. Pymol[57] and the Vienna-PTM webserver[58] were used to prepare and manipulate PDB files.

Two perturbation paths were used for lysine methylation and serine phosphorylation, where the hydrogen atoms were allowed or not allowed to perturb into heavy atoms. Additionally, serine was perturbed into phosphoserine of total net charge of -1 and -2. Each perturbation process in water was simulated in 21 steps with equidistant λ-points and 5 ns per point, while the bound state was simulated in 41 steps with equidistant λ-points and three independent 8 ns long

simulation per point. Three independent sets of simulations of the bound state were generated by randomly assigning the initial velocities using a different random seed.

Analysis of generated free-energy data

Several methods were used to compute the free energy change upon studied alchemical processes, including multistate Bennett acceptance ratio (MBAR),[59] Bennett acceptance ratio (BAR),[15] extended thermodynamic integration with MBAR and linear interpolation for predicting the ensemble average of the partial derivatives of the Hamiltonian with respect to λ at λ-points that were not simulated (101 and 81 equidistant λ-points for peptides in water and the bound state, respectively) (exTI$_{MBAR}$ and exTI$_{lin}$)[14] and thermodynamic integration (TI).[13]

A number of additional properties were calculated from obtained data, including error estimators from MBAR (ΔG_{err}^{MBAR}), BAR (ΔG_{err}^{BAR}), error estimates from bootstrapping the data for a 100 times and calculating the standard deviation of the obtained distribution ($\Delta G_{BSerr}^{method}$, where *method* stands for any of the methods mentioned above: MBAR, BAR, exTI$_{MBAR}$, exTI$_{lin}$ and TI), overlap integrals (OI) and differences in exTI predictions from neighboring λ-points (ΔG_{err}^{exTI}), as well as fractions of simulated times for which the computed free energy difference is within a given tolerance cutoff from the value obtained using the full simulation time $\Delta\Delta G_{frac}^{MBAR\ tol}$. Considered tolerances were 1 kJ mol^{-1}, ½ kT, kT and 1 kcal mol^{-1}, while fractions of data and OI were ranging from 0.1 – 0.9 with a step of 0.1.

Simulation update scheme

Several update schemes for simulation time at different λ-points were tested. Starting from initial set of simulations (6 or 11 equidistant λ-points with 0.5 or 1 ns per point), convergence of different λ-segments was tested. In particular, segments of 0.025, 0.05, 0.1 or 0.2 in terms of $\Delta\lambda$ were considered. The convergence criteria were based on 4 tolerance cutoffs applied to the error

estimate from MBAR and / or the computed free energy difference using half of the simulation data ($\Delta\Delta G_{half}^{MBAR\ tol}$, equivalent to $\Delta\Delta G_{frac}^{MBAR\ tol}$ with the fraction of 0.5). All tested schemes were separated in two main update strategies. Update strategy 1 consists of iterations in which all the λ-segments are checked for the convergence. For segments (between λ_1 and λ_2) that do not reach the criteria yet, the simulation time at λ_1 and λ_2 are doubled by prolonging the simulations. Simultaneously a new λ point is introduced at $\lambda = \frac{\lambda_1 + \lambda_2}{2}$, and two independent simulations are performed for the same amount of simulation time. This approach ensures that the wall time of each simulation stays constant (allowing for efficient parallelization on a high performance computing (HPC) cluster) and that in each iteration the simulation time at already simulated λ-points doubles, while at newly added it immediately reaches the same amount.

In the update strategy 2, the total amount of simulated time (the same as in the initial iteration) is distributed over the λ-segment which did not reach the convergence (the same criteria as in the strategy 1), based on the weights obtained from one or a combination of all four calculated properties: error estimate from MBAR; computed free energy difference using half of the simulated data and MBAR; overlap integral; and differences in exTI predictions from neighboring λ-points. This approach ensures that the total simulation time in each iteration stays constant, facilitating plannability of the use of computer recourses.

RESULTS AND DISCUSISON

Perturbation topology builder

An initial step in perturbation free-energy calculations is the definition of the alchemical pathway. A dual topology is widely used to transform all atoms of one compound from real atoms into non-interacting dummy atoms, and the other way around for the other compound. This approach is arguably inferior to matching the common atoms of the two compounds and

perturbing the rest, especially in case of compounds with a high level of similarity. However, this approach usually involves a cumbersome manual procedure of defining the matches and mismatches. To address this, an automated tool able to generate a perturbation topology of two compounds by finding their maximum common substructure was developed. It is based on the VF algorithm for graph isomorphism matching,[41] where the potential common substructures are enumerated iteratively. A pruning function ensures reasonable running times, even though this is not a guarantee, as graph isomorphism matching is of exponential complexity. This notwithstanding, several tests on small molecules, post-translational modifications and amino-acid mutations were completed within seconds. Importantly, while enumerating the substructures, the algorithm also evaluates the perturbations based on molecular topologies, making it possible to guide the search towards the maximum number of matched atom types as defined by the force field used. This is arguably one of the most common choices, and when performed on lysine trimethylation modification (Figure 1A) results in all atoms being mapped to each other, with three hydrogen atoms assigned for perturbation into methyl groups (note that methyl groups within the GROMOS force field are modeled as a single united-atom particle). On the other hand, one can design other matching criteria, for instance by minimizing the number of perturbed bonds, which would lead to a different perturbation topology. In the case of lysine methylation the hydrogen atoms are perturbed into dummy atoms, while the methyl groups are grown from dummy particles, with the rest of the atoms being matches (Figure 1B). It is worth noting that the algorithm allows for wide flexibility in tuning the MCS search by setting different penalty weights for different types of individual perturbations compared to each other, including atoms types, perturbed bonds, angles, improper dihedrals or dihedrals. The enumerated solutions are sorted according to the score defined by the penalty weights. In addition, this feature can be

used not only to select the preference towards a specific type of perturbation, but also to generate different perturbation definitions (pathways) between a given pair of compounds of interest, which can be tested for their performance.

When it comes to matching ring structures, the algorithm allows for two options: 1) partial match of polycyclic compounds where only a complete match of individual rings is allowed (Figure S1); and 2) only complete match is allowed. In the case of matching between ring and non-ring atoms, three options are provided: 1) partial match of maximum of 2 atoms (that share a bond); 2) partial match of only one atom; and 3) no match of a ring to a non-ring atom is allowed (Figure S2). Option 1 was chosen to be default in both cases, as it permits for matching larger maximum common substructure. Note that allowing for a partial match of three or more atoms in a ring structure would potentially affect the sampling of the conformational space of the end states. Arguably, allowing a partial match of two atoms that share a bond would not have such an effect, however this assumption remains to be tested in simulation. For this reason, the alternative choices are provided allowing one or no atom as a partial match.

When it comes to multiple topologies (3 or more), the algorithm is able to find the expected common substructures when applied on a set of simple compounds, including alkane chains and cycloalkanes of same length with additional methyl groups at different position (Figure S3 and 4) or polycycles as shown in Figure 2. In addition to these simplified test cases, in a recently published work,[44] EDS topologies were generated using this tool based on automatically recognized scaffold among different sets of molecules. Note that an EDS topology is a single topology (defining reference state Hamiltonian) that can represent multiple molecules by switching atom types, where the free energy differences between the molecules are calculated using a one step perturbation approach from the reference state. They include a set of 16

glutamate receptor A2 (GRA2) allosteric modulators, a set of 8 trypsin inhibitors and a set of 10 phenylethanolamine N-methyltransferase inhibitors. Such a multi-topology approach can also be used to generate a closed thermodynamic cycle for a set of compounds. Note however, finding maximum common substructure on a set of multiple topologies requires longer runtimes (in minutes), compared to dual topologies.

While it is tempting to try applying this multi-topology algorithm on a big set of diverse compounds, e.g. screening libraries of compounds, this would most probably lead to intractable running times. However, such sets of compounds are arguably also not relevant in the context of the EDS methodology or evaluation of the cycle closure, since the number of states / perturbation lags would be too large for meaningful calculations. On the other hand, this toolkit offers an alternative way of tackling a large set of compounds by employing pairwise generation of dual perturbation topologies, potentially in combination with utilizing EDS techniques[42–44] on small subgroups of similar compounds or by optimizing the choices of pairwise perturbations as proposed by Liu et al.[60]

Application of the tool on a set of histone-related post-translational modifications and convergence analysis of the perturbation calculations

Lysine acetylation, lysine 3-methylation and serine phosphorylation, as representative modifications of each type of the most important histone modifications were chosen as a test set. Firstly, the influence of the type of the perturbation pathway as a possible source of inaccuracy and potential convergence issues was tested. To this end, the predicted values of the partial derivatives of the Hamiltonian with respect to λ (at λ-points that were not simulated) were used. For example, such predictions from different simulated λ-points largely diverge for lysine methylation when perturbing hydrogen into a methyl group and a related bond (Figure 1A and

Figure 3A), which is not the case for the same modification and a different perturbation path (Figure 1B and Figure 3B). Similar discrepancies of $\partial H/\partial\lambda$ predictions, but to a much smaller extent, were observed for serine phosphorylation involving perturbations of hydrogen to heavy atoms with related bonds, leading to the conclusion that such type of perturbation paths are to be avoided. As observed in previous studies,[61,62] the source of this behavior is arguably due to bond perturbation, since even small changes in the distance between bonded atoms lead to a great difference in the potential energy or its derivative with respect to the coupling parameter λ. Therefore, all other results reported here do not involve bond perturbation.

To test how different choices of simulated λ-points affect the computational costs, i.e. total simulation time required to reach converged calculations, a number of different scenarios were assumed, with different number of initial λ-points and allowing or prohibiting non-equidistant λ points and different simulation time. Four convergence criteria of different stringency were considered, and the amount of simulation time required to reproduce the reference data (full simulation data with MBAR) within 1 kJ mol^{-1}, ½ kT, kT and 1 kcal mol^{-1} was determined. Note that the convergence criteria were applied to each λ-segment to avoid compensation of errors. In addition, the performance of different methods for computing the free energy differences from simulation data was explored. Allowing non-equidistant λ-points drastically decreases computational costs of free energy calculations, while additional improvement is observed with variable simulation time for different λ-points (Figure 4). This analysis also revealed that MBAR and extended TI in combination with MBAR for predicting the $\partial H/\partial\lambda$ at non-simulated λ-points (exTIM$_{BAR}$) perform similarly and better than the other tested methods, followed by BAR and exTI$_{lin}$. Interestingly, plain thermodynamic integration did not reach convergence for any of the tested systems of tolerance cutoff applied, with the largest one of 1 kcal mol^{-1}. One should note

that the minimum possible simulation time (Figure 4), as well as an ideal set of λ-points (Figure 4 most right bar) can be determined only in a post-analysis using extensive simulation data. Obviously, one would prefer to select the optimal protocol *a priori*.

To address this, the attention was turned to a number of error estimators and other related properties (Table S1), including error estimates from MBAR and BAR, error estimates from bootstrapping, overlap integrals and differences in exTI predictions from neighboring λ-points, as well as fractions of simulated times for which the computed free energy difference is within a given tolerance cutoff from the complete simulation time. The simulated data was separated in different subsets of λ-points and simulated times by applying non-overlapping and overlapping sliding windows on the time and λ scale, respectively. In this way, a large number of independent simulation data subsets was generated to improve statistical analyses.

While displaying various forms and levels of correspondence to the true errors, $\Delta\Delta G_{\text{true}}$, estimated as the difference to the free energy computed from the complete extensive simulation data, such error estimates fail at providing quantitative predictions of true errors. For examples see Figure 5. This result, in addition to potential limitations of the estimators themselves, is most probably the consequence of sampling issues. Most notably, it is impossible to predict the effect of simulation diffusing to other energy minima in the phase space that were not visited thus far in the simulation. This in turn highlights the importance of adequate sampling and enhanced sampling techniques.[11,44,63,64] Interestingly, making pairwise comparisons between these different estimators shows that some pairs correlate almost perfectly with each other, while some show anti-correlation or no correlation (Figure S5 and 6). This suggests that the examined estimators rely and report on different aspects of convergence behavior of the simulations.

Furthermore, to explored how the convergence of calculations depend on the set of error estimators directly measurable from the simulations, accuracy-sensitivity analysis was performed. Here, the accuracy is defined as the fraction of considered segments (simulation subsets fulfilling a given criterion with respect to the error estimators) yielding the free energy difference within a given tolerance cutoff from the full simulation data (i.e. $\Delta\Delta G_{\text{true}}$ smaller than the tolerance). The sensitivity on the other hand was calculated as the fraction of accurate segments identified by the error estimator compared to the total number of accurate segments, up to a given tolerance cutoff. In other words, high accuracy implies low number of inaccurate segments, while high sensitivity reflects low number of accurate segments that are not detected. This analysis shows that by applying a well-defined cutoff to a given estimator, one can achieve a high accuracy, however, often to a price of relatively low sensitivity (Figure 6, S7). For example, when using MBAR as free-energy estimator and simulations in which ΔG_{err}^{MBAR} is smaller than 1 kJ mol^{-1}, one can expect that roughly 75 % of the segments are accurate within 1 kT, with 95 % of all accurate segments detected. This suggests, as expected, that the estimators, although lacking a quantitative agreement, provide at least a qualitative match to the true errors in the free-energy differences. Taken together with the observations from the pairwise correlation analysis (low correlation between some pairs of estimators as seen in Figure S6), one can potentially create a collective variable, i.e., a combination of the studied estimators, to improve the quantitative aspect of the predictions. To test this, different combinations and conditions applied to different estimators were screened. In particular, $\Delta\Delta G_{frac}^{MBAR\ tol}$ (fractions ranging 0.1 – 0.9 in steps of 0.1), ΔG_{err}^{MBAR}, ΔG_{BSerr}^{MBAR}, ΔG_{err}^{exTI} with the tolerance of 1 kJ mol^{-1}, ½ kT, kT and 1 kcal mol^{-1}, and OI in the range 0.1 – 0.9 in steps of 0.1 were used. In Figure 7, the

individual points represent sets of segments that fulfil the above defined combinations and conditions.

When applied to the combined data from all simulated GGXGG systems, both high accuracy and selectivity is achieved (Figure 7, top). More than 60 % of tested cases provide an accuracy higher than 0.9 for a stringent 1 kJ mol^{-1} tolerance, with some of them showing a sensitivity greater than 0.4. This behavior is even more prominent for less stringent tolerance cutoffs, where a sensitivity greater than 0.8 can be achieved with an accuracy of 0.9 within the tolerance of 1 kcal mol^{-1}. However, this approach unfortunately breaks down when applied to the individual systems. In particular, for lysine methylation and acetylation perturbation the overall well-functioning conditions retain high accuracy and selectivity (Figure 7, second and third row), however, fail to repeat the success for the serine phosphorylation perturbation, dropping in the percentage of conditions for which the accuracy is higher than 0.9 (Figure 7, last two rows). This highlights that the development of such a protocol is strongly system-dependent and would be applicable only to systems with very similar perturbation pathway, which is often not the case. Potentially, this assumption could hold for instance when applying the same perturbation between two ligands in water and the bound state, a setup often applied in relative binding free energy calculations.[1,26,35,36]

Iterative simulation update scheme for free energy calculations

With the observed limitations revealed by the above analyses, an alternative approach of determining the choice and simulation times at different λ-points, based on an iterative simulation update scheme, was further developed and tested. Three design principles were used: 1) optimization of the usage of available computational resources, 2) flexibility and 3) practical usage. In particular, an initial set of simulation time and λ-points is selected, e.g., 11 equidistant

λ-points with 1 ns long simulation each. The update scheme is then run for a number of cycles in such a way that in each iteration, the user-defined sets of λ-segments are tested for convergence using a set of conditions with a given tolerance cutoff. For each segment (e.g. $0.1 < \lambda < 0.2$) that did not reach the convergence, simulations at its bordering λ-points ($\lambda = 0.1$ and $\lambda = 0.2$) get prolonged to reach double simulated time, while an extra intermediate λ-point is added ($\lambda = 0.15$) and two independent simulations are run starting from configurations obtained in simulations of each of the bordering points. This design (update strategy 1) ensures that at each iteration (cycle) the amount of simulation is quadrupled at the non-converged segments, while the simulation time of individual simulations remains constant, increasing the parallelization potential. Importantly, this update strategy, regardless of the system, initial set of λ-points and simulation times, convergence conditions or λ-segments checked for convergence, reached a very low discrepancy with the true free energy difference, with only four instances of the discrepancy exceeding kT, out of which two exceeded 1 kcal mol^{-1} (Figure 8). It is worth noting, that a complex system of lysine methylation in the context of a histone tail bound to a PHD finger domain was added to the test set. The perturbation simulations for that system were performed using the GROMACS simulation package, as its performance allows for more rapid data collection. Similarly to other analyses in this study, lysine acetylation and methylation in water reach the convergence relatively quickly. Serine phosphorylation and lysine methylation within a complex environment go through the maximum number of iterations without reaching the convergence, however, still approach the true free energy difference to a relative low discrepancy below kT for almost all tested conditions.

To test whether a further optimization in effective usage of computer resources could be achieved, an alternative flavor of this iterative simulation update strategy to the same set of

systems was applied. This procedure (update strategy 2) in each iteration cycle distributes a total amount of simulation time (set by the user) over a set of λ-segments identified as not converged, using the same types of user-defined criteria as in the strategy 1. The simulation length of each individual λ-point simulated within an iteration (cycle) is determined using weighting based on error estimators. Interestingly, regardless of the choice of four such estimators (ΔG_{err}^{MBAR}, $\Delta\Delta G_{half}^{MBAR}$, overlap integral and ΔG_{err}^{extTI}) as well as a combination thereof, almost identical results were obtained (data not shown). This strategy indeed lowers the total simulation times compared to the strategy 1, however also with a cost of increasing the discrepancy with the true free energy difference (Figure 9), with six instances exceeding a relatively high cutoff of 1 kcal mol^{-1}.

Strikingly, in both update strategies all such cases of increased discrepancy to the true free energy, i.e. in cases that failed to reproduce the true free energy difference within 1 *kT*, are based on schemes using either a lower number of initial λ-points ($\Delta\lambda^{init}$ = 0.2) or shorter initial simulations (simulation time of 0.5 ns), or a combination of both. It is also worth noting that these schemes employ exclusively the free-energy difference on half of the simulated data ($\Delta\Delta G_{half}^{MBAR}$) as the convergence criterium. On the other hand, when 1-ns-long simulations at 11 equidistant λ-points are applied in the initial iteration and/or the error estimate from MBAR (ΔG_{err}^{MBAR}) as the convergence criterium, the discrepancy with the true-free energy difference remains well below *kT* regardless of the system, update strategy and other criteria used (Figure 8 and 9). Furthermore, this update approach is able to identify the amount of simulation time (and λ-points) needed to reach the convergence, which is strongly depending on the system in question. More importantly, this approach provides great flexibility in terms of the setup of the simulation scheme, which can be adjusted according to the complexity of the perturbation system

or availability of computational resources. This design also allows for performing additional cycles (iterations) of simulations if needed. Note that the implementation of the update scheme is independent of the simulation package, as it takes the simulated data as input and returns a set of λ-points and simulation times as output in each iteration.

CONCLUSIONS

In this study two important challenges related to the perturbation free energy approaches were addressed. Firstly, an automated tool for generating dual perturbation topologies (GROMOS and GROMACS file formats) based on a maximum common substructure algorithm is introduced. In each enumeration step of generating matched subgraphs, force-field defined topology parameters are checked and stored. This allows for a flexible maximum common substructure search by setting weighted preference towards minimizing perturbations of atom types or different types of bonded interactions such as bonds, angles, improper dihedrals or dihedrals. In addition to perturbation topologies between two states, this algorithm is able to generate a combined perturbation topology for a set of multiple topologies (three or more), which is primarily aimed to be used in combination with EDS techniques, but can also be applied to define closed thermodynamic cycles.

Secondly, this study investigates how the efficiency and the convergence of perturbation free energy calculations can be improved by optimizing the choice of λ-points and simulation times, using a set of test systems based on common post-translational modifications accruing in histone tails. Allowing non-equidistant λ-points and non-constant simulation time reduced drastically the total simulation time required to reach the convergence in tested systems. This is not surprising, however also not practically useful, as such an optimization is only possible in post-processing when the converged result is already known. To address this, the attention was turned to

examining how a number different properties (e.g. bootstrapping error estimate, MBAR error estimate, overlap integral, etc.) calculated from the simulated data predict the level of convergence. While a qualitative agreement to the difference from the true ΔG (the free energy change calculated from an extensive simulation) was observed, all tested estimators failed at quantitative prediction of the error. As some pairs of the estimators showed no correlation to each other, in order to improve the predictive power, defining new estimators as collective variables (combinations) of the individual ones was attempted, however with a limited success only. Building further on the performed analyses, an iterative simulation update scheme was proposed, aimed at effective usage of computational resources by prioritizing simulations at λ-points with poor convergence. In particular, in each iteration (cycle) of the update scheme a new set of λ-points and related simulation times is suggested based on thus far collected simulation data with two main update strategies available, oriented at utilizing the parallelization potential of a computer cluster or at further reducing the total amount of simulation time. Various flavors of the update scheme were successfully applied on test systems with different complexity of perturbation. Importantly, different levels of enhanced efficiency was achieved through allowing both variable simulation times and non-equidistant λ-points, convergence within 1 kT discrepancy from the true ΔG was reached in almost all cases, especially when more simulation time is collected in the initial step.

Finally, it can be expected that the advances presented in this study might significantly improve the applicability of perturbation free energy methodology in different contexts ranging from estimation of protein stability to binding affinity calculations to rational drug development. To this end, this toolkit is provided as an open source python package via a github repository (https://github.com/drazen-petrov/SMArt).
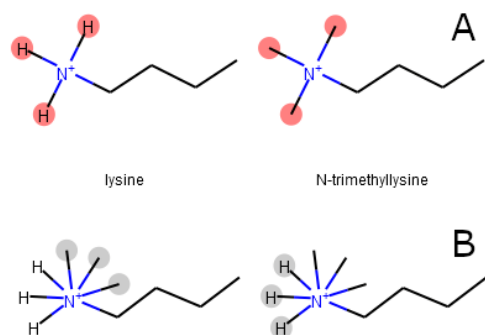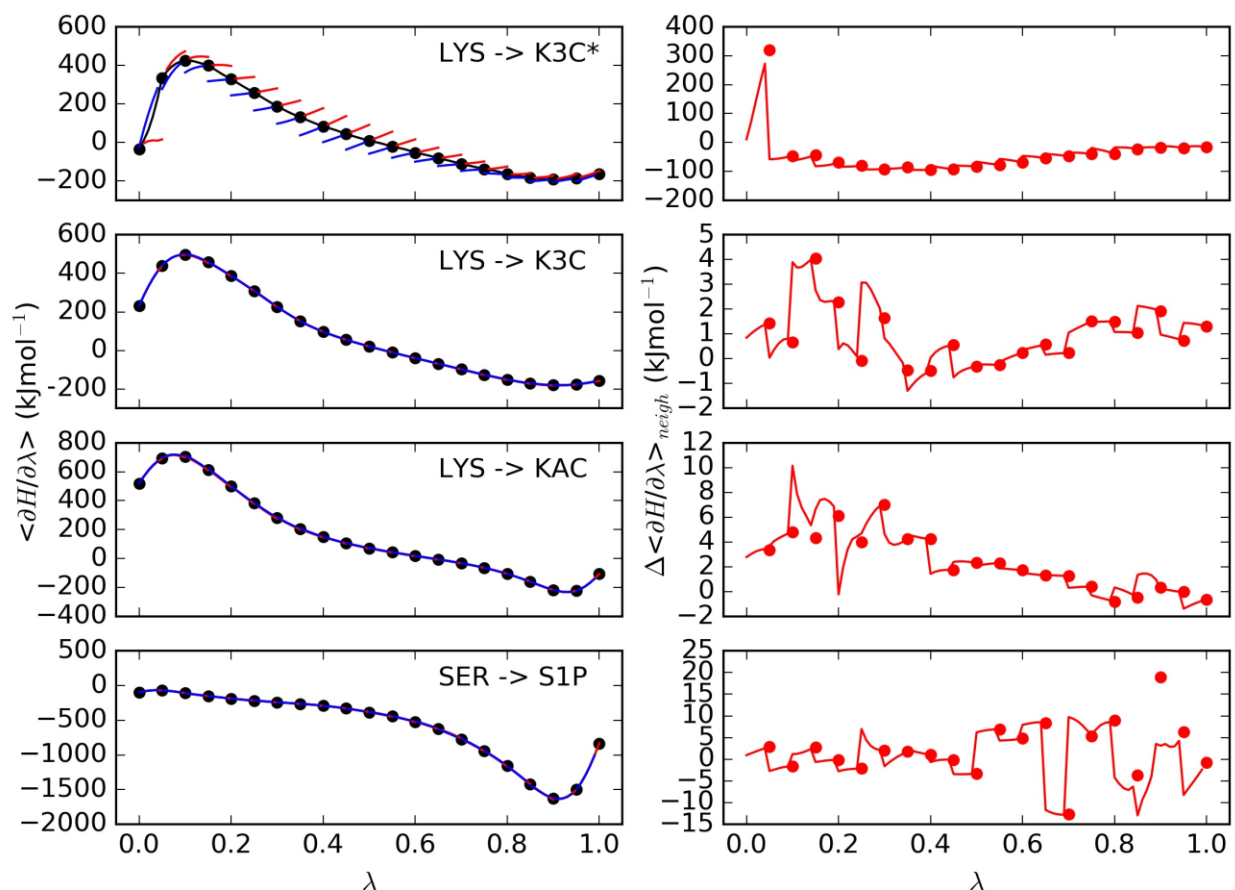
FIGURES

Figure 1. Alternative scenarios for the methylation of the lysine sidechain. Maximum common substructure of lysine (left) and its methylated form (right). A) The perturbation topologies are generated by maximizing the number of matching atoms and B) by minimizing the number of perturbed bonds (B). Perturbed atoms are highlighted in red, while dummy atoms are highlighted in gray.
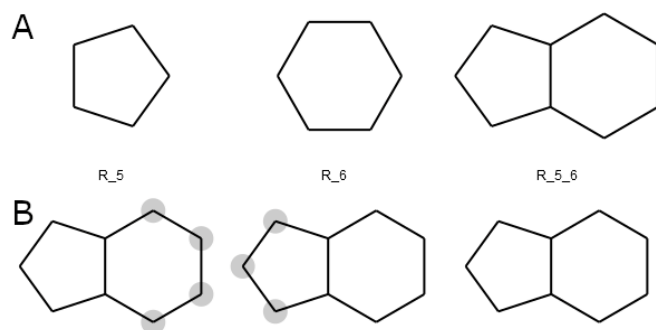
Figure 2. Multi-topology perturbation of three ring compounds. Top row represents individual compounds, while the bottom row corresponding EDS states where dummy atoms are highlighted in gray.

Figure 3. Representative $\partial H/\partial \lambda$ curves. Simulated (black points) and predicted (using extended TI approach) $\partial H/\partial \lambda$ values are shown on the left side of the figure. The predictions to the smaller λ-points (left from the simulated point) are shown with short blue lines and the predictions to the larger λ-points (right from the simulated point) are shown with short red lines. The black solid line represents the linear interpolation between the predicted values of $\partial H/\partial \lambda$. Right side of the figure shows the total difference in the predictions (between the blue and the red lines of the left panel) from neighboring simulated λ-points as red lines. At the simulated λ-points these differences were calculated between the simulated values (black points left) and the predictions

from the closest larger λ-point. Red points represent the difference between the simulated values (black points left) and the predictions from the closest smaller λ-point. LYS -> K3C* and LYS -> K3C present lysine methylation perturbation involving and excluding the perturbation of hydrogens to methyl groups and related bonds, respectively. Note that the predicted and linearly interpolated values can only be visually distinguished for the LYS -> K3C* (top row left) perturbation, whereas they simply overlap in the other cases.
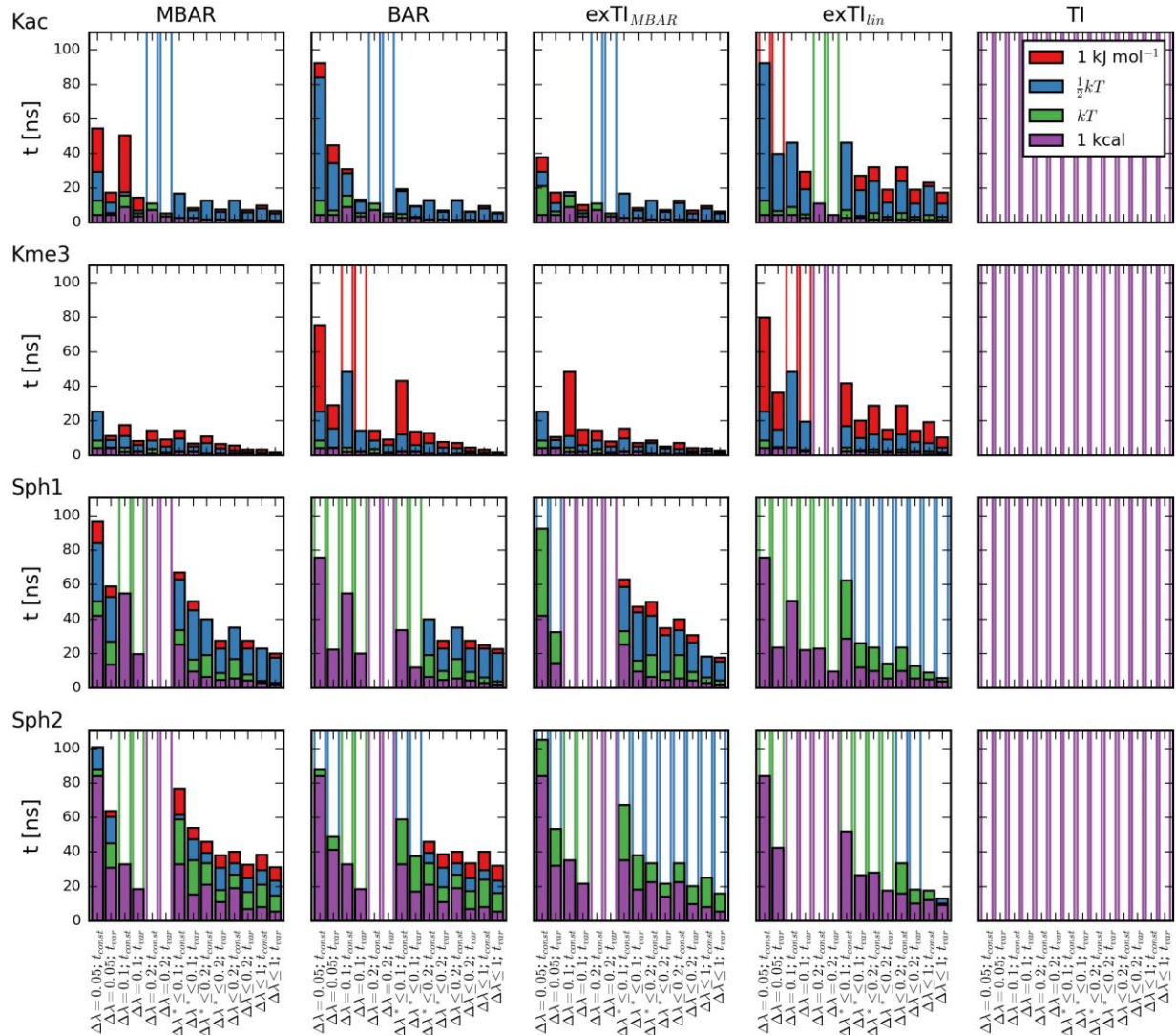
Figure 4. Minimal simulation time needed to obtain the free energy difference within the tolerance cutoff from the free energy calculated using the complete extensive set of simulation data (21 λ points with 5 ns each, using MBAR). MBAR and extended TI with MBAR (exTI$_{MBAR}$) used for predicting $\partial H/\partial \lambda$ values converge to the free-energy difference obtained from complete data similarly and faster than BAR, that converges faster than extended TI in combination with linear interpolation (exTI$_{lin}$), while plain thermodynamic integration (TI) fails to reach convergence regardless of the perturbation in question or tolerance cutoff applied. Each bar represents aggregate data over all simulations with indicated restrictions on the choice of the

simulated λ-points (equidistant or non-equidistant) and simulated time per point (constant or variable). For example, the left-most bar represents 21 equidistant λ-points ($\Delta\lambda = 0.05$) with constant simulation time per point (t_{const}), while the right-most bar represents the ideal case, i.e. any choice of λ-points ($\Delta\lambda \leq 1$, non-equidistant points with the maximum distance of 1) with variable simulation time (t_{var}). $\Delta\lambda^*$ represents an additional restriction to the choice of the λ-points, such $\Delta\lambda^* \leq 0.1$ and $\Delta\lambda^* \leq 0.2$ must include the points from the points from the sets of equidistant λ-points $\Delta\lambda = 0.1$ and $\Delta\lambda = 0.2$, respectively. Note that all sets (bars) include the end states ($\lambda = 0$ and $\lambda = 1$). Bars are colored according to the tolerance cutoff of 1 kJ mol^{-1}, ½ kT, kT and 1 kcal mol^{-1}, in red, blue, green and purple respectively. Empty bars with edge colors represent that no convergence for the stated restrictions was reached. Note that a calculation was considered converged only if each λ-subsegment reached convergence (according to the defined tolerance), such that potential compensation of error is avoided. Free energy differences are computed from the simulation data using 5 different methods.
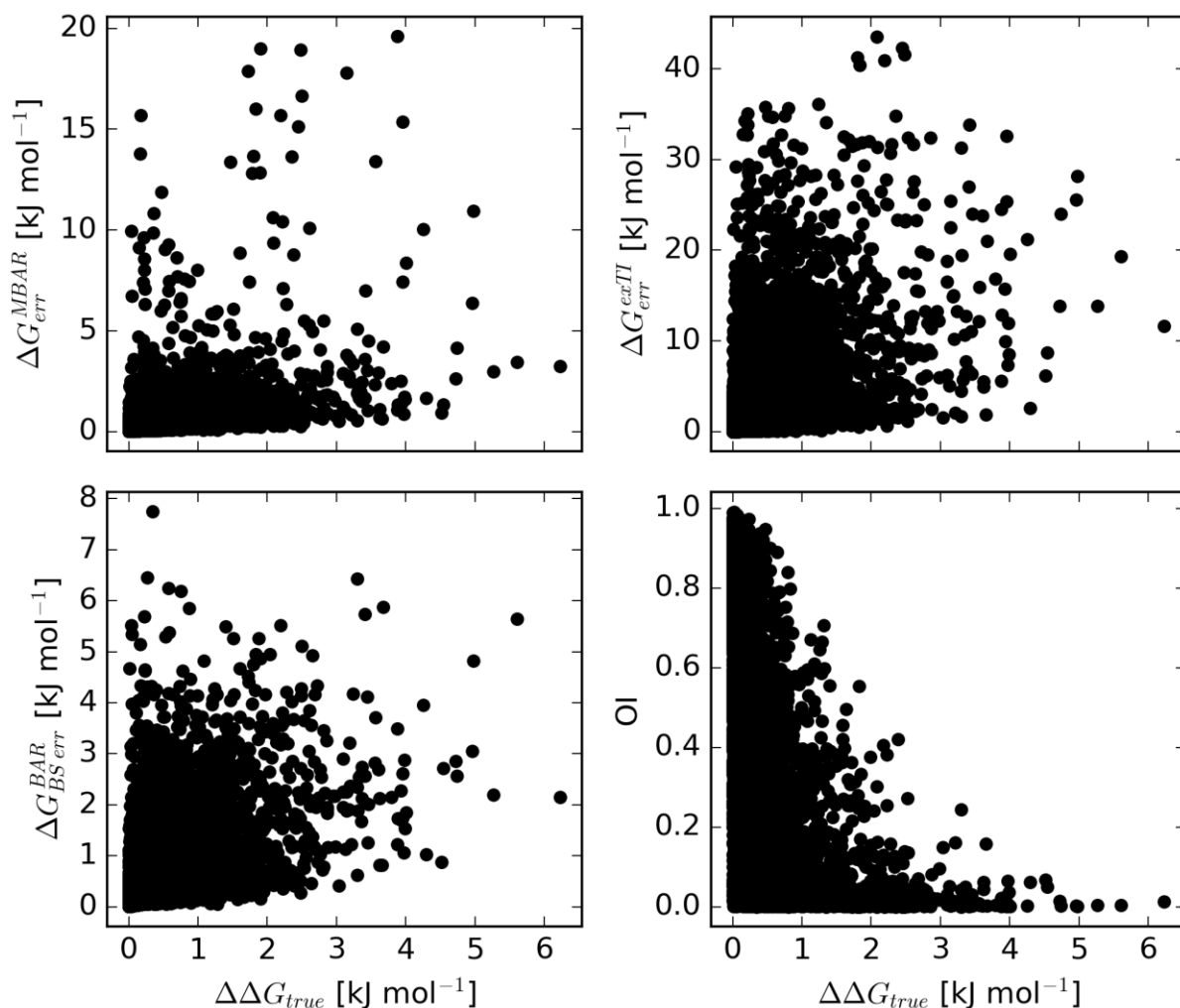
Figure 5. Error estimators compared to the true error $\Delta\Delta G_{true}$, calculated as the difference to the free energy from the complete extensive set of simulation data (21 λ points with 5 ns each, using MBAR). While a certain relationship of the estimators to the true errors is observed, generally a low level of correlation indicates low predictive power. Example estimators shown, including the error estimate from MBAR, bootstrapping error estimate from BAR, the differences in predicted $\partial H/\partial \lambda$ from exTI and overlap integrals (OI) for the neighboring λ-points. More than 8000 segments (simulation subsets) of aggregated data for different simulation times, sets of simulated λ points and simulated peptide systems is shown.
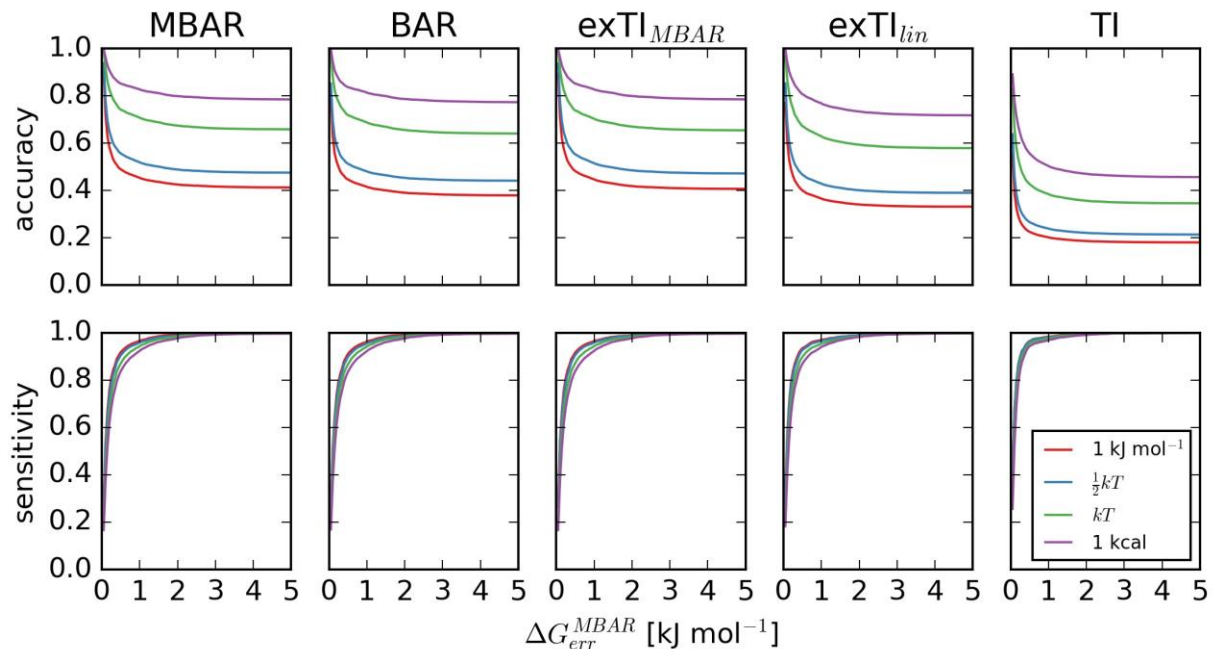
Figure 6. Accuracy and sensitivity analysis of calculated ΔG values using error estimate from MBAR. Even though the correlation between this error estimator and the true error is rather weak, at low values of ΔG_{err}^{MBAR} high accuracy can be achieved. However, with the high accuracy, low sensitivity was observed, which is defined as a fraction of accurate segments in the total number of accurate segments. Average over all aggregated data for different simulation times, sets of simulated λ points and simulated peptide systems are shown. Lines colored according to the tolerance cutoff applied of 1 kJ mol^{-1}, ½ *kT, kT* and 1 kcal mol^{-1}, in red, blue, green and purple respectively.
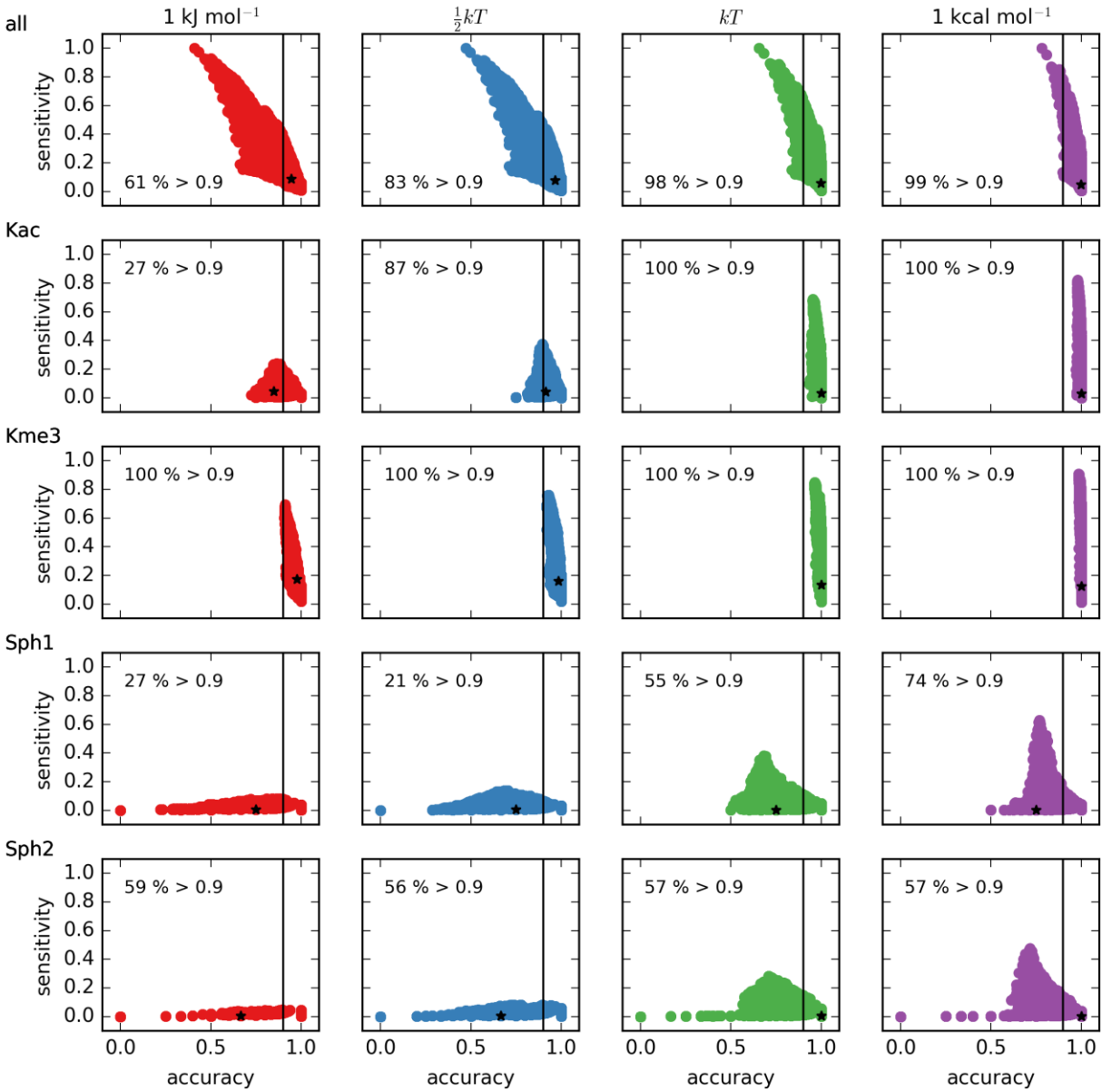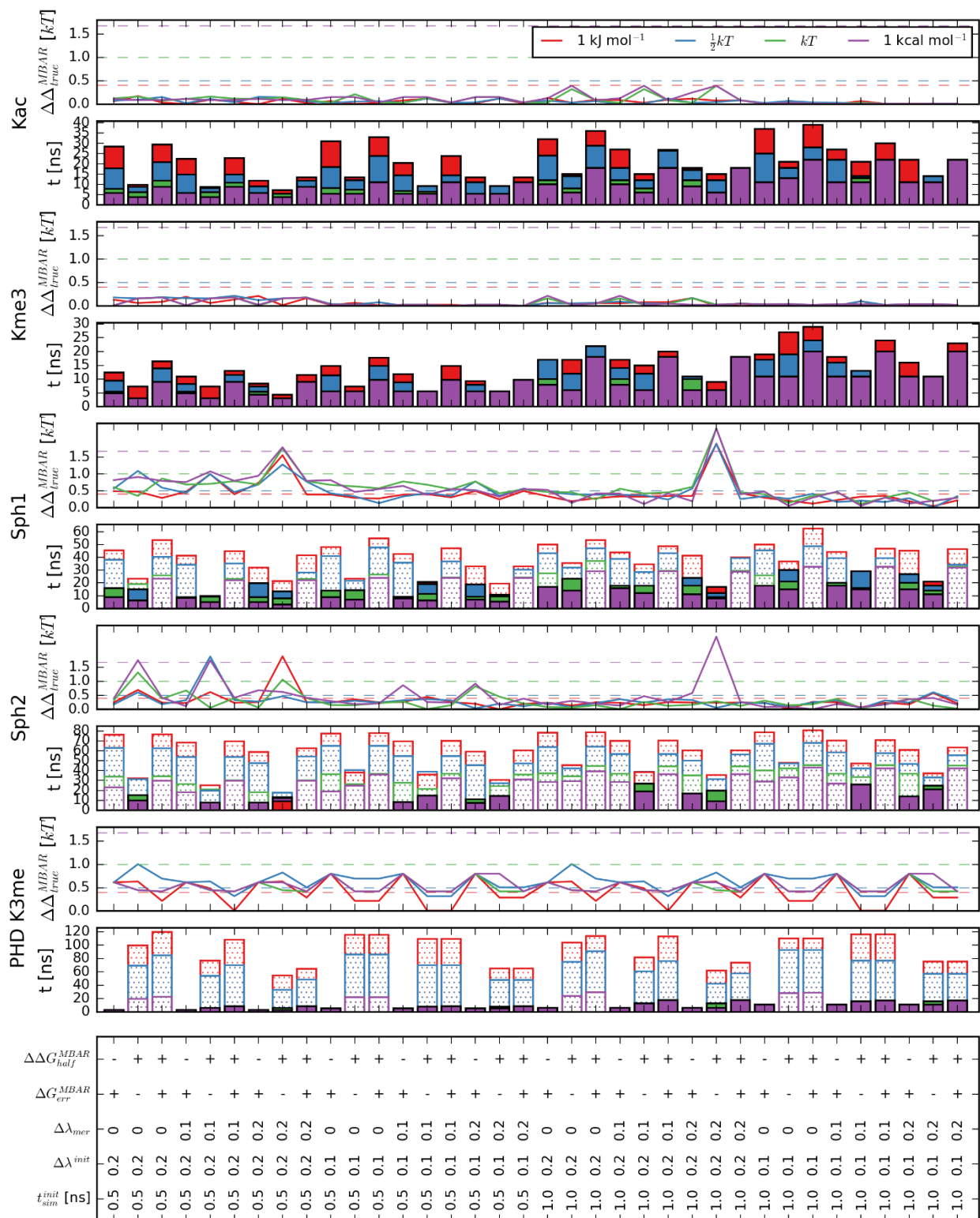
Figure 7. Accuracy and sensitivity analysis of a collective variable by simultaneously applying several error estimators in an attempt to optimize both accuracy and sensitivity (top) and testing such an approach on individual systems. Screening for a set of conditions applied to the error estimators, including the error estimate from MBAR, the differences in predicted $\partial H/\partial \lambda$ from exTI, overlap integrals (OI) for the neighboring λ-points, and the difference in the calculated free energy from half and total simulated data, shows that a large fraction of applied conditions on the

aggregated data from all simulated GGXGG systems provides high accuracy of more than 0.9, with even more than 60 % of such for a stringent tolerance cutoff of 1 kJ mol^{-1} or staggering 99 % for a weak cutoff of 1 kcal mol^{-1} (top). However, this approach breaks down when utilized for individual systems (last 4 rows), where the sets of optimized conditions perform well for methylation and acetylation perturbation providing the accuracy of at least 0.9 up to 1 kJ mol^{-1} tolerance for all instances, while poorly in the case of phosphorylation modifications providing the same level of accuracy with only about 55 % of tested sets of conditions. Each point represents a subset of data according to the applied conditions to the error estimators with the accuracy and sensitivity calculated with a given tolerance cutoff. For example, one such combination $\Delta G_{err}^{MBAR} < 1$ kJ mol^{-1}; $\Delta G_{BSerr}^{MBAR} < \frac{1}{2}\, kT$; $\Delta G_{err}^{exTI} < kT$; $\Delta\Delta G_{half}^{MBAR} < 1$ kcal mol^{-1}; OI > 0.1 is depicted as a black star in each panel. Note that the last 4 rows contain only the corresponding data from the top row that falls in the group of conditions with the accuracy higher than 0.9. Points colored according to the tolerance cutoff applied: 1 kJ mol^{-1}, $\frac{1}{2}\, kT$, kT and 1 kcal mol^{-1}, in red, blue, green and purple respectively.

Figure 8. Simulation time, convergence of tested simulation update schemes (strategy 1), together with the accuracy, represented as the difference from the free energy calculated using the complete extensive simulated data (21 λ-points with 5 ns each for peptide simulations and 41 λ-points and 8 ns in three replicates each for the bound system using MBAR). Very good level of accuracy is reached almost regardless of the scheme applied or test system, with only a few exceptions reaching discrepancy greater than 1 kT from the true free energy difference, obtained from the complete simulated data. A broad range of total simulation times was reached, from only a few ns, up to more than 150 for some update schemes for the bounded state. Even though it is hard to postulate rules from this data, general trend shows that considering longer λ-segments for convergence leads to improved efficiency without losing the accuracy, while collecting more simulated data in the initial cycle improves accuracy. Each bar represents an applied update scheme determined by the initial simulation time and λ-points, criteria to reach convergence, as well as the size of considered λ-segments for convergence. Filled bars indicate reached convergence, while empty ones with a dotted pattern indicate that the maximum number of update iterations has been reached (such that the maximum simulation time of a single λ-point does not exceed the total simulated time), colored according to the tolerance cutoff of 1 kJ mol^{-1}, ½ kT, kT and 1 kcal mol^{-1}, in red, blue, green and purple respectively (lower panels). Each line in higher panels represents the difference from the true value calculated from the complete set of simulations.

Figure 9. Simulation time, convergence of tested simulation update schemes (strategy 2), together with the accuracy, represented as the difference from the free energy calculated using the complete extensive simulated data (21 λ-points with 5 ns each for peptide simulations and 41 λ-points and 8 ns in three replicates each for the bound system using MBAR). Similar trends in the data were observed as compared to the update strategy 1 (Figure 8), however with improved efficiency and somewhat worse accuracy. Very good level of accuracy is reached almost regardless of the scheme applied or test system, with only a few exceptions reaching discrepancy greater than 1 kT from the true free energy difference, obtained from the complete simulated data. A broad range of total simulation times was reached, from only a few ns, up to more than 120 ns for some update schemes for the bounded state. Each bar represents an applied update scheme determined by the initial simulation time and λ-points, criteria to reach convergence, as well as the size of considered λ-segments for convergence. Filled bars indicate reached convergence, while empty ones with a dotted pattern indicate that the maximum number of update iterations has been reached (such that the maximum simulation time of a single λ-point does not exceed the total simulated time), colored according to the tolerance cutoff of 1 kJ mol^{-1}, ½ kT, kT and 1 kcal mol^{-1}, in red, blue, green and purple respectively (lower panels). Each line in higher panels represents the difference from the true value calculated from the complete set of simulations.

TABLES.

Table 1. Post-translational modifications for which perturbation free-energy calculations were performed.

codes		modification
LYS	Kac	lysine acetylation
LYS	Kme3	lysine trimethylation
SER	Sph1	serine phosphorylation (net charge -1)
SER	Sph2	serine phosphorylation (net charge -2)

ASSOCIATED CONTENT

Supporting Information.

List of error estimators; MCS of partial ring match; MCS of ring non-ring match; Multi-state EDS topology for a set of alkane chains; Multi-state EDS topology for a set of rings; Correlation between several error estimators; Heatmap of correlation coefficients between all pairs of calculated error estimators; Accuracy and sensitivity analysis (PDF)

AUTHOR INFORMATION

Corresponding Author

*Drazen Petrov

drazen.petrov@boku.ac.at

Author Contributions

DP designed the study, performed the simulations and analyses, and wrote the manuscript.

ABBREVIATIONS

MD, molecular dynamics; TI, thermodynamic integration, exTI, extended thermodynamic integration; MBAR, Multistate Bennett Acceptance Ratio; BAR, Bennett Acceptance Ratio; HPC, high performance computing; MCS, maximum common substructure

REFERENCES

(1)   Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Alchemical Free Energy Calculations: Ready for Prime Time? In *Annual Reports in Computational Chemistry, Vol 3*; Spellmeyer, D. C., Wheeler, R. A., Eds.; Annual Reports in Computational Chemistry; Elsevier: Amsterdam, Netherlands, 2007; Vol. 3, pp 41–59.

(2)   Seeliger, D.; de Groot, B. L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, *98* (10), 2309–2316.

(3)   Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical Free Energy Methods for Drug Discovery: Progress and Challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150–160.

(4)   de Ruiter, A.; Oostenbrink, C. Efficient and Accurate Free Energy Calculations on Trypsin Inhibitors. *J. Chem. Theory Comput.* **2012**, *8*, 3686–3695.

(5)   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.

(6)   Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chem. Int. Ed Engl.* **2016**, *55* (26), 7364–7368.

(7)   Petrov, D.; Daura, X.; Zagrovic, B. Effect of Oxidative Damage on the Stability and Dimerization of Superoxide Dismutase 1. *Biophys. J.* **2016**, *110* (7), 1499–1509.

(8)   Malde, A. K.; Stroet, M.; Caron, B.; Visscher, K. M.; Mark, A. E. Predicting the Prevalence of Alternative Warfarin Tautomers in Solution. *J. Chem. Theory Comput.* **2018**, *14* (8), 4405–4415.

(9)   Petrov, D.; Tunega, D.; Gerzabek, M. H.; Oostenbrink, C. Molecular Modelling of Sorption Processes of a Range of Diverse Small Organic Molecules in Leonardite Humic Acid. *Eur. J. Soil Sci.* **2019**, *508*, 276.

(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.

(11) de Ruiter, A.; Oostenbrink, C. Advances in the Calculation of Binding Free Energies. *Curr. Opin. Struct. Biol.* **2020**, *61*, 207–212.

(12) Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; van Vlijmen, H.; Tresadern, G.; de Groot, B. L. Large Scale Relative Protein Ligand Binding Affinities Using Non-Equilibrium Alchemy. *Chem. Sci.* **2020**, *11* (4), 1140–1152.

(13) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300–313.

(14) Ruiter, A. de; Oostenbrink, C. Extended Thermodynamic Integration: Efficient Prediction of Lambda Derivatives at Nonsimulated Points. *J. Chem. Theory Comput.* **2016**, *12* (9), 4476–4486.

(15) Bennett, C. H. Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245–268.

(16) Goette, M.; Grubmüller, H. Accuracy and Convergence of Free Energy Differences Calculated from Nonequilibrium Switching Processes. *J. Comput. Chem.* **2009**, *30* (3), 447–456.

(17) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78* (14), 2690–2693.

(18) 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.

(19) Bruckner, S.; Boresch, S. Efficiency of Alchemical Free Energy Simulations. I. A Practical Comparison of the Exponential Formula, Thermodynamic Integration, and Bennett's Acceptance Ratio Method. *J. Comput. Chem.* **2011**, *32*, 1303–1319.

(20) Bruckner, S.; Boresch, S. Efficiency of Alchemical Free Energy Simulations. II. Improvements for Thermodynamic Integration. *J. Comput. Chem.* **2011**, *32*, 1320–1333.

(21) 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.

(22) Pham, T. T.; Shirts, M. R. Optimal Pairwise and Non-Pairwise Alchemical Pathways for Free Energy Calculations of Molecular Transformation in Solution Phase. *J. Chem. Phys.* **2012**, *136* (12), 124120.

(23) Lee, T.-S.; Lin, Z.; Allen, B. K.; Lin, C.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Sherman, W.; York, D. M. Improved Alchemical Free Energy Calculations with Optimized Smoothstep Softcore Potentials. *J. Chem. Theory Comput.* **2020**, *16* (9), 5512–5525.

(24) König, G.; Glaser, N.; Schroeder, B.; Kubincová, A.; Hünenberger, P. H.; Riniker, S. An Alternative to Conventional λ-Intermediate States in Alchemical Free Energy Calculations: λ-Enveloping Distribution Sampling. *J. Chem. Inf. Model.* **2020**, *60* (11), 5407–5423.

(25) Reinhardt, M.; Grubmüller, H. Determining Free-Energy Differences Through Variationally Derived Intermediates. *J. Chem. Theory Comput.* **2020**, *16* (6), 3504–3512.

(26) de Ruiter, A.; Petrov, D.; Oostenbrink, C. Optimization of Alchemical Pathways Using Extended Thermodynamic Integration. *J. Chem. Theory Comput.* **2021**, *17* (1), 56–65.

(27) Homeyer, N.; Gohlke, H. FEW: A Workflow Tool for Free Energy Calculations of Ligand Binding. *J. Comput. Chem.* **2013**, *34* (11), 965–973.

(28) Christ, C. D.; Fox, T. Accuracy Assessment and Automation of Free Energy Calculations for Drug Design. *J. Chem. Inf. Model.* **2014**, *54* (1), 108–120.

(29) 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* (5), 348–354.

(30) Klimovich, P. V.; Mobley, D. L. A Python Tool to Set up Relative Free Energy Calculations in GROMACS. *J. Comput. Aided Mol. Des.* **2015**, *29* (11), 1007–1014.

(31) Loeffler, H. H.; Michel, J.; Woods, C. FESetup: Automating Setup for Alchemical Free Energy Simulations. *J. Chem. Inf. Model.* **2015**, *55* (12), 2485–2490.

(32) Fu, H.; Gumbart, J. C.; Chen, H.; Shao, X.; Cai, W.; Chipot, C. BFEE: A User-Friendly Graphical Interface Facilitating Absolute Binding Free-Energy Calculations. *J. Chem. Inf. Model.* **2018**, *58* (3), 556–560.

(33) Jespers, W.; Esguerra, M.; Åqvist, J.; Gutiérrez-de-Terán, H. QligFEP: An Automated Workflow for Small Molecule Free Energy Calculations in Q. *J. Cheminform.* **2019**, *11* (1), 26.

(34) 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.

(35) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D., Jr; 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.

(36) Kim, S.; Oshima, H.; Zhang, H.; Kern, N. R.; Re, S.; Lee, J.; Roux, B.; Sugita, Y.; Jiang, W.; Im, W. CHARMM-GUI Free Energy Calculator for Absolute and Relative Ligand Solvation and Binding Free Energy Simulations. *J. Chem. Theory Comput.* **2020**, *16* (11), 7207–7218.

(37) Strahl, B. D.; Allis, C. D. The Language of Covalent Histone Modifications. *Nature* **2000**, *403*, 41–45.

(38) Jenuwein, T.; Allis, C. D. Translating the Histone Code. *Science* **2001**, *293*, 1074–1080.

(39) Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **2007**, *128*, 693–705.

(40) Zhang, X.; Zhao, D.; Xiong, X.; He, Z.; Li, H. Multifaceted Histone H3 Methylation and Phosphorylation Readout by the Plant Homeodomain Finger of Human Nuclear Antigen Sp100C. *J. Biol. Chem.* **2016**, *291* (24), 12786–12798.

(41) Cordella, L. P.; Foggia, P.; Sansone, C.; Vento, M. Subgraph Transformations for the Inexact Matching of Attributed Relational Graphs. *Graph Based Representations in Pattern Recognition* **1998**, *12*, 43–52.

(42) Christ, C. D.; van Gunsteren, W. F. Enveloping Distribution Sampling: A Method to Calculate Free Energy Differences from a Single Simulation. *J. Chem. Phys.* **2007**, *126* (18), 184110.

(43) Sidler, D.; Cristòfol-Clough, M.; Riniker, S. Efficient Round-Trip Time Optimization for Replica-Exchange Enveloping Distribution Sampling (RE-EDS). *J. Chem. Theory Comput.* **2017**, *13* (6), 3020–3030.

(44) Perthold, J. W.; Petrov, D.; Oostenbrink, C. Toward Automated Free Energy Calculation with Accelerated Enveloping Distribution Sampling (A-EDS). *J. Chem. Inf. Model.* **2020**, *60* (11), 5395–5406.

(45) Schmid, N.; Christ, C. D.; Christen, M.; Eichenberger, A. P.; van Gunsteren, W. F. Architecture, Implementation and Parallelisation of the GROMOS Software for Biomolecular Simulation. *Comput. Phys. Commun.* **2012**, *183*, 890–903.

(46) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015/9**, *1–2*, 19–25.

(47) Reif, M. M.; Hünenberger, P. H.; Oostenbrink, C. New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 3705–3723.

(48) Petrov, D.; Margreitter, C.; Grandits, M.; Oostenbrink, C.; Zagrovic, B. A Systematic Framework for Molecular Dynamics Simulations of Protein Post-Translational Modifications. *PLoS Comput. Biol.* **2013**, *9* (7), e1003154.

(49) Margreitter, C.; Reif, M. M.; Oostenbrink, C. Update on Phosphate and Charged Post-Translationally Modified Amino Acid Parameters in the GROMOS Force Field. *J. Comput. Chem.* **2017**, *38* (10), 714–720.

(50) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, 1981; pp 331–342.

(51) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(52) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* **1995**, *102* (13), 5451–5459.

(53) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

(54) Páll, S.; Hess, B. A Flexible Algorithm for Calculating Pair Interactions on SIMD Architectures. *Comput. Phys. Commun.* **2013**, *184* (12), 2641–2650.

(55) Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(56) Beutler, T. C.; Mark, A. E.; Vanschaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free-Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(57) Schrodinger, L. The PyMOL Molecular Graphics System, Version 1.3r1. 2010.

(58) Margreitter, C.; Petrov, D.; Zagrovic, B. Vienna-PTM Web Server: A Toolkit for MD Simulations of Protein Post-Translational Modifications. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W422-6.

(59) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129* (12), 124105.

(60) Liu, S.; Wu, Y.; Lin, T.; Abel, R.; Redmann, J. P.; Summa, C. M.; Jaber, V. R.; Lim, N. M.; Mobley, D. L. Lead Optimization Mapper: Automating Free Energy Calculations for Lead Optimization. *J. Comput. Aided Mol. Des.* **2013**, *27* (9), 755–770.

(61) Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B* **2010**, *114* (32), 10235–10253.

(62) König, G.; Brooks, B. R.; Thiel, W.; York, D. M. On the Convergence of Multi-Scale Free Energy Simulations. *Mol. Simul.* **2018**, *44* (13–14), 1062–1081.

(63) 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.

(64) Jiang, W.; Thirman, J.; Jo, S.; Roux, B. Reduced Free Energy Perturbation/Hamiltonian Replica Exchange Molecular Dynamics Method with Unbiased Alchemical Thermodynamic Axis. *J. Phys. Chem. B* **2018**, *122* (41), 9435–9442.