# Best Practices for Multiple Alchemical Free Energy Calculations [Article v0.1]

**Antonia S. J. S. Mey[1*], Bryce K. Allen[2], Hannah E. Bruce Macdonald[3], John D. Chodera[3*], David F. Hahn[9], Maximilian Kuhn[1,10], Julien Michel[1], David L. Mobley[4*], Levi N. Naden[5], Samarjeet Prasad[6], Andrea Rizzi[2,7], Jenke Scheen[1], Michael R. Shirts[8*], Gary Tresadern[9], Huafeng Xu[2]**

[1]EaStCHEM School of Chemistry, David Brewster Road, Joseph Black Building, The King's Buildings, Edinburgh, EH9 3FJ, UK; [2]Silicon Therapeutics, Boston, MA, USA; [3]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York NY, USA; [4]Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine, Irvine, USA; [5]Molecular Sciences Software Institute, Blacksburg VA, USA; [6]National Institutes of Health, Bethesda, MD, USA; [7]Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY, USA; [8]University of Colorado Boulder, Boulder, CO, USA; [9]Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30, Beerse B-2340, Belgium; [10]Cresset, Cambridgeshire, UK

FC: @Volunteer update initial contributor list to reflect sections copied from part 1.

*This LiveCoMS document is maintained online on GitHub at https://github.com/alchemistry/alchemical-best-practices; to provide feedback, suggestions, or help improve it, please visit the GitHub repository and participate via the issue tracker.*

*This version dated October 10, 2024*

## Abstract

FC: The original best practices guide will be split into two documents: the first will gives best practices for running a single calculation (and give details on theory/ system-specific setup), while the second will give best practices for running multiple calculations (with a focus on automation). The first will become version 2.0 of the current guide, while the second will be submitted as a new manuscript.
This document outlines the contents of the second document for multiple calculations. This will also have a focus on methods for automation.

FC: @Volunteer Update abstract to highlight the split of the document and focus on multiple calculations.

Alchemical free energy calculations are a useful tool for predicting free energy differences associated with the transfer of molecules from one environment to another. The hallmark of these methods is the use of "bridging" potential energy functions representing *alchemical* intermediate states that cannot exist as real chemical species. The data collected from these bridging alchemical thermodynamic states allows the efficient computation of transfer free energies (or differences in transfer free energies) with orders of magnitude less simulation time than simulating the transfer process directly. While these methods are highly flexible, care must be taken in avoiding common pitfalls to ensure that computed free energy differences can be robust and reproducible for the chosen force field, and that appropriate corrections are included to permit direct comparison with experimental data.

In this paper, we review current best practices for several popular application domains of alchemical free energy calculations performed with equilibrium simulations, in particular relative and absolute small molecule binding free energy calculations to biomolecular targets.

Received: 5 August 2020
Accepted: 25 November 2020

https://doi.org/10.33011/livecoms.2.1.18378

## Contents

## Todo list

FC: The original best practices guide will be split into two documents: the first will gives best practices for running a single calculation (and give details on theory/ system-specific setup), while the second will give best practices for running multiple calculations (with a focus on automation). The first will become version 2.0 of the current guide, while the second will be submitted as a new manuscript.

This document outlines the contents of the second document for multiple calculations. This will also have a focus on methods for automation. 1section*.2

## 1 What are alchemical free energy methods?

FC: @Volunteer to add section briefly summarising the same section in the first paper, directing readers to the first paper, and highlighting their use a high-throughput, multiple calculation setting.

## 2 Prerequisites and Scope

FC: @Volunteer to update clearly delineating the separation between the first (single calculation) and second (multiple caluclations) papers.

FC: @Volunteer to a) integrate the sections cut directly from v1 of the paper and b) consider adding further sections as required.

## 3 How should alchemical simulations be applied to drug discovery?

FC: Is this section too broad? Should we split accross rest of paper?

FC: Should we clearly seperate and discuss needs of lead opt. and virtual screening? For example recent work with fast ABFE from Sandbox AQ.

### 3.1 Is the expected accuracy of the computation sufficient?

FC: Move section within document

The requisite level of accuracy is another important consideration. If the goal is to guide lead optimization when many compounds will be synthesized, free energy calculations can be appealing even with accuracies in the 1–2 kcal/mol range [1], but if the number of compounds to be synthesized is very small, this accuracy may not be enough to provide much value.

Here we provide a simple estimate of the value provided by alchemical free energy calculations in lead optimization. Let $P(\Delta\Delta G)$ be the probability distribution of the changes in the binding free energies of a new set of molecules during one round of lead optimization, and let $P(\Delta\Delta G^{\dagger}\,|\,\Delta\Delta G)$ be the conditional probability of the binding free energy change computed by the free energy calculations, $\Delta\Delta G^{\dagger}$, given the actual change $\Delta\Delta G$. The latter conditional probability can be modeled by a normal distribution

$$P(\Delta\Delta G^{\dagger}\,|\,\Delta\Delta G) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\Delta\Delta G^{\dagger} - \Delta\Delta G)^2}{2\sigma^2}\right), \quad (1)$$

where $\sigma$ signifies the accuracy of free energy calculations. Here we assume that there is no systematic bias in the free energy calculations, i.e., on average, the free energy change computed by free energy calculations agrees with the actual free energy change. Additional analysis of this type is presented in Brown et al. [2]

In lead optimization guided by free energy calculations, we will likely only synthesize and experimentally test molecules that are predicted to have favorable free energy changes. We are thus interested in how often a molecule predicted to bind stronger actually turns out to bind stronger. In other words, we are interested in the conditional probability:

$$P(\Delta\Delta G < 0\,|\,\Delta\Delta G^{\dagger} < 0). \quad (2)$$

For illustrative purposes, consider a proposed set of new molecules, and assume that the changes proposed in these molecules yield a set of relative binding free energies that follow a normal distribution. That is, assume that the standard deviation in the relative binding free energies for the changes represented is $RT \ln 5$ (corresponding to a 5-fold change in the binding affinities), and that 1 in 10 new molecules have increased binding affinity ($\Delta\Delta G \leq 0$). Under such assumptions, the conditional probability in Eq. 2 can be easily computed.

If the accuracy of a collection of free energy calculations is $\sigma = 1$ kcal/mol, $P(\Delta\Delta G < 0\,|\,\Delta\Delta G^{\dagger} < 0) = 0.35$, which means that out of every 10 molecules selected for predicted favorable free energy change, on average 3.5 molecules will have actual favorable free energy change. In other words, selection by free energy calculations yields 3.5 times more molecules of improved affinities than selection without free energy calculations under these assumptions.

Available computational resources and timescales of motion also factor into this initial analysis. An individual free energy calculation involves simulations at many different intermediate states (perhaps 20-40 or more) and each of these must typically be long enough to capture the relevant motions in the system. If such motions are microsecond events or longer, the computational cost of running 20-40 microsecond or longer simulations for each of $N$ ligands will likely be prohibitive for most users with today's hardware. On the other hand, if key motions are fast and minimal (as is often assumed in practice), much shorter simulations may be sufficient.

### 3.2 Can I afford the calculation?

Furthermore, are available computational resources sufficient that throughput will be reasonable compared to needs of experimental collaborators working on this system? How many ligands (*N*) can you afford to handle given your computational resources? As cloud computing becomes more available, in-house GPU clusters may not be necessary if calculations are not run on a regular basis. This analysis should be done up front as part of "counting the cost" of involvement in a particular project. In some cases, the analysis may conclude that free energy calculations will not be feasible for the proposed problem. Here, by "cost", we refer not just to financial cost of the calculations relative to experiments, but also time – can the calculations be run faster than experiments are done? How will the relevant resource and opportunity costs factor in? Both computation and experiment require human time, supplies (of different sorts), and equipment. In the extreme limit, for example, it would not make sense to spend a month running a binding free energy calculation if the equivalent experiment could be done in a day with resources already on hand. Such issues should be considered before deciding to conduct binding free energy calculations.

### 3.3 Is an exploratory study what I want?

An additional consideration is how much is known about your particular target, ligand binding modes in the target, and any relevant motions – essentially, has it been studied enough to know whether it might be suitable for free energy calculations? It is important to know if the system has hardly been studied, because should the initial calculations perform poorly, the effort may turn into an attempt to understand the relevant sampling, force field, or system preparation problems.

If you are unsure whether your project is feasible, as mentioned above, one recommended option is to conduct a short exploratory study to assess tractability for a small number of ligands. This can be sufficient to get an initial idea of feasibility and accuracy of the calculations for the proposed target [3].

FC: @FC Move entire section to second paper.

Many practitioners expect alchemical methods to provide valuable guidance for drug discovery, and to exhibit accuracy superior to most alternative approaches for suitable targets [4]. Successful application in industry may require considerable knowledge of the "domain of applicability" of free energy calculations – where they work well and where they will not [5]. Successful application also requires robust protocols for preparing, submitting and analysing alchemical calculations. In this regard, the issues mentioned in the previous section such as understanding the suitability and timescales to capture the structure activity relationships (SAR), and performing up-front tests of performance are all relevant to drug discovery applications. Without venturing too far into details of system setup, which is beyond the scope of this article, we highlight some critical factors affecting accuracy and successful application.

### 3.4 Capturing experimental conditions

The calculations aim to capture the alchemical change from one ligand to another as accurately as possible. Therefore, it is necessary to consider details of the experimental setup, such as pH. Biological assays are usually run at neutral pH but this is not always the case. For example, some enzymes exhibit pH-dependent activity and assays may thus be done in conditions other than neutral pH. Therefore, computational protein and ligand preparation protocols should reflect experimental pH.

The formal charge and/or tautomeric state of the small molecules can change within a series of analogs, necessitating care in treatment. Additionally, medicinal chemistry efforts might deliberately modify the pKa of a series to modify drug properties, requiring explicit efforts to incorporate these changes into alchemical calculations.

To ensure modeling matches experiment, we also need to accurately prepare and simulate the same system – which requires understanding what protein construct is used in the bioassay. For instance, does the X-ray structure that is to be used for the calculations match the construct used for screening (i.e. only the catalytic domain vs. full length, monomer vs. dimer, etc.) [6]? Also, were certain co-factors or partner proteins required in the bioassay?

### 3.5 Is my binding mode accurate?

FC: @volunteer Seems like this section is partly duplicated (e.g. "Are you prepared to deal with any binding mode challenges?" in part 1). @Volunteer to tidy this up, moving appropriate information to first article and focussing here on challenges in a high-throughput setting, e.g. should I use multiple binding poses from docking? How should I select them?

As also mentioned, good performance of alchemical calculations requires an accurate representation of the ligand binding mode, usually from a high quality X-ray crystal structure. If more than one structure is available, the modeler should pay attention to choose the most suitable. The quality of the structure can be a concern, and the reader is referred to work of Warren et al. for a detailed discussion of choosing optimal structures for structure-based modeling [7].

It is also useful to study the structure activity relationship and understand the expected impact of any mutations on the binding site, such as whether side chain movement in the protein will be required, and whether there is evidence of this in any alternative X-ray structures of the same protein. Often, only one protein and water configuration is used for a series of alchemical calculations, so this needs to be capable of accommodating the smallest through to largest ligands in a way that allows stable and well behaved simulations. This can provide a practical limit on the alchemical changes that are feasible, though a simple work-around can be to separate compounds into sub-series for different calculations.

If multiple structures are available there is some evidence the higher affinity complex can give better match to experiment [8], at least in some cases. However, ligands and proteins can also undergo unexpected changes in binding mode for related ligands, which can make these issues more complex to deal with [9].

### 3.6 Input setup and scale of calculations

In a drug discovery setting it is normal to consider dozens (or more) of ligands and it is necessary to align them in the binding site. There is no detailed study of how different alignment approaches may affect results, but the user should be aware of some practical considerations. Tools are available to compare the ligands and build the combined topologies that define the changes between one ligand and another [10–12]. In simple terms, providing poor alignment to these tools will make this job harder. Docking with restraints is often beneficial in this regard. Particularly, fixing the 3D spatial position of the scaffold using maximal common substructure (MCSS) restrained docking can help provide well aligned input for the topology generation. Nevertheless, in this case careful attention is still needed to ensure consistency of alignment for identical substituents. Another alternative is to manually edit the same core and add/modify the changing substituents. This provides assurances that coordinates for the non-perturbed portion of the structure remain identical and aromatic substituents, for instance, have consistent dihedral angles. However, it is not feasible for many compounds and therefore automation is desirable.

Finally, the role of water in ligand binding is not always well understood and it can be crucial to capture the changes in binding site solvation during ligand binding. Can crystallographic waters be retained? Do they clash with some of the larger ligands used in the alchemical perturbation? See Sec. **??** for different strategies that can be applied to dealing with waters. Generally, before launching large numbers of alchemical free energy calculations it is always recommended to test the system using classical MD simulations and limited numbers of alchemical perturbations. Metrics such as ligand and protein RMSD and RMSF can be inspected, along with visual inspection of simulations, to ensure the system is stable and likely to be suitable for alchemical calculations.

Running binding free energy calculations in a drug discovery application will typically require the use of software or tools to facilitate the large number of calculations. Commercial implementations such as FEP+, OpenEye Tools, or Flare allow for a fast setup and deployment to GPU hardware in minutes, but may have limited ability to customize calculations [13, 14]. Commercial tools can be expensive in some cases, but non-commercial tools are becoming more straight forward to use to run alchemical free energy calculations [10–12, 14–17].

For relative free energy calculations, various graph topologies or maps of calculations are possible, and choices may depend on the target application. For instance, if the goal is to accurately assess the relative binding energy of a small number of compounds, possibly with challenging syntheses, the map of perturbations should contain as many connections between compounds as affordable. However, when running calculations on hundreds of compounds a so called *star-map* (see Fig. 1**A**) can be used that just contains one connection per compound: perturbing every compound to a central ligand, typically the crystal structure ligand [18]. In this way the top-ranking examples can be readily identified and submitted to additional calculations in a second round. Alternatively, if the goal is to achieve the smallest possible error with minimal computational expense, certain graph topologies provide benefits [19, 20]

### 3.7 Making predictions, understanding errors

For prospective drug discovery applications there are several other considerations including understanding likely errors and taking selection bias into account.

It is crucial when proposing compounds for synthesis to have some idea of the underlying error or uncertainty in the predictions. A retrospective assessment can give an indication of prospective performance for similar molecules [21]. Beyond this, several parameters provide useful indicators of performance. For example error estimates provided by free energy estimators that are too large can highlight poorly converged simulations [8]. Hysteresis, either within cycles in the perturbation network or between forward and backward perturbations can be checked [22] to indicate problematic perturbations involved in cycles connecting many compounds (See also Secs. **??** and **??**). Once synthesis and testing of compounds is complete a standard strategy is to look back at how the calculations performed. In this regard it is important to consider the issue of selection bias upfront. It is tempting to only synthesize the compounds predicted to be most active, thus a narrow range of calculated activity is tested that imposes limits on the statistical assessment of performance, ideally example molecules from across the range of predicted activity can be assessed or corrections can be applied based on previous recommendations [23]. For a more detailed discussion on checking the robustness of your alchemical free energy calculation see also Sec. **??**.

In summary, the successful use of alchemical calculations, particularly for drug discovery, requires working in the domain of applicability, using a high quality X-ray structure of the target bound to compounds in the series, and testing the approach retrospectively to ensure the system setup is well-behaved. Always assess your confidence in the resulting predictions and communicate this when discussing with experimentalists. Consider performing repeat calculations for at least some of the perturbations in the study. There are many accounts of success of alchemical calculations, the methods show good performance towards the goal of binding free energy prediction. However, it is important to have realistic expectations.

Structure based drug design projects are often capable of improving potency relatively quickly, even with only limited application of computational approaches and the range of activity narrows to just two-to-three log units. It may seem hard to have impact with substantially different, more potent, stand-out compounds in this scenario, but binding free energy predictions can still be extremely useful for ensuring activity is maintained as other properties are optimized. An interesting cost benefit analysis has shown the value of activity prediction, see discussion above and articles such as [1]. From a drug discovery point of view, alchemical calculations are expanding their domain of applicability, and there are reports of success using homology models [24] and GPCRs [25, 26] for instance, as well as enabling charge change and scaffold hopping [27, 28], but these systems are undoubtedly more difficult. In the meantime, use cases are expanding to resistance prediction, selectivity prediction , solubility prediction – an exciting future for alchemical calculations [29–31].

## 4 Which calculations should I run, and when?

FC: Possibly add discussion of methods which are not easily characterised as single/ multi e.g. lambda dynamics? Maybe a better distinction is calculations where we are interested in more than two end states.

FC: @OpenFE team to add discussion of how to run and adaptively update perturbation networks.

FC: Possibly add discussion of different protocols, especially active learning. Best practices may not be well established but many companies are running at scale.

FC: Discussion of merits of ABFE/ RBFE for virtual screening? See @Gary Tresadern comment from first meeting - lots of discussion around ABFE calculations for virtual screening.

### Perturbation maps

FC: @Jenke Scheen/ Mary Pitman/ OpenFE team to update discussion of perturbation maps?

Based on the input ligand series, a perturbation map or network can be planned. Recent heuristics have shown the more connected the perturbation network the better. However, there is a way to optimize network structure while minimizing the number of perturbations that need to be computed reducing the resulting computational cost [19, 20]. Sometimes the introduction of intermediates that are not part of the original congeneric series are essential to avoid ring breaking, or to deal with perturbations that would otherwise result in large numbers of atoms being inserted or deleted. Some commercial tools have good underlying heuristics but may fail with complicated input, needing user validation in particular when dealing with chiral compounds.

In some cases, during the lead optimization stage, or for very large datasets that would benefit from rougher initial free energy ranking, or in cases where perturbations would be rather large, a star shaped network as seen in Fig. 1 **A** is used. However, adding redundancy into the network means that a better error analysis can be carried out by looking at cycle closure errors as discussed in sec. **??**, with an example given in Fig. 1**B**.

Methods from experimental design have been applied to the construction of the perturbation maps. Yang et al. [19] showed how to optimize the perturbation map by selecting a fixed number of calculations from the pairwise perturbations so that the resulting set of calculations minimize the total variance. Xu [20] showed how to optimize the perturbation map by allocating different amounts of simulation time to different pairwise perturbations so as to minimize the total variance, given the total simulation time of all the perturbation calculations. Both approaches lead to substantial reduction in the statistical error of the estimated free energies.

https://doi.org/10.33011/livecoms.2.1.18378

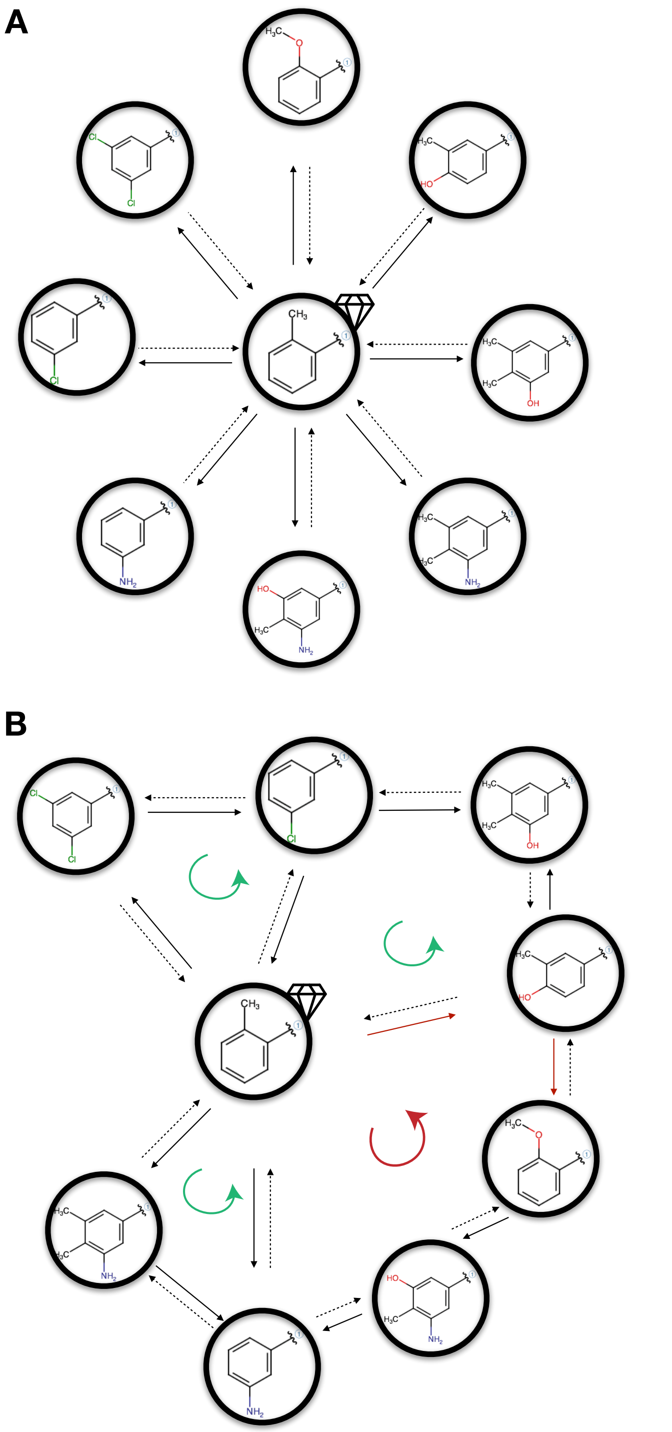


**Figure 1. Examples of perturbation networks** (**A**) Star shaped network with the crystal structure in the center. (**B**) Network with cycle closures (see more on this in Sec. **??**). Arrows indicate the direction of the perturbation. Fully converged binding free energy calculations yield binding free energy changes which sum to zero around any closed cycle. However, in practice errors may not sum to zero around closed cycles, providing a way to look for potential sampling problems. Here in (B), green cycles indicate cycles with hypothetically good cycle closure, red those with poor cycle closure. The red arrow indicates a poorly converged simulation that would give rise to bad cycle closures. The diamond indicates the use of a crystallographic binding mode.

## 5 How should I automate my calculations within my workflow?

FC: Include e.g. automated selection of restraints for ABFE, automated selection of where to concentrate sampling time e.g. paper from Huafeng Xu, discussion of available tools?

FC: @OpenFE team well suited. Discuss tools like alchemiscale? Discuss e.g. integration with REINVENT/ Maize for active learning?

## 6 Data analysis

### 6.1 Uncertainty estimation

FC: @Peter Coveney/ @Agastya P Bhati to cover uncertainty estimation for multiple simulations. Move cycle closure etc sections here.

#### Cycle closure error

Relative free energy calculations, which compute the change in free energy on making a change to a molecule (e.g. adding a functional group to a ligand) may provide an additional opportunity for error/consistency checking. Particularly, such calculations are often done to span a graph or tree of free energy calculations [20, 22, 32]. In some cases the free energy change to go between molecules A and B can be obtained via multiple transformation pathways. This allows a type of consistency checking where we assess how much the free energy change for that transformation in practice differs from equivalence.

Significant deviations of agreement from the same transformation by different routes typically indicate insufficient configurational sampling along the lambda schedule of one or more of the transformations involved. This approach may be generalised to sets of connected transformations given the requirement that the sum of free energy changes along edges of a closed cycle should be zero. This analysis is called "cycle closure". In practice, such thermodynamic cycles do not actually sum to zero, and deviations become increasingly large as the size of the cycle increases owing to propagation of error. Though no firm guidelines have emerged, it may be judicious to perform additional configurational sampling along edges of a network that are involved in cycles closing poorly. This may be done by extending the duration of simulations, or by averaging free energy changes over multiple repeats. The latter approach may yield more reproducible free energy changes, but at the expense of a stronger bias on the estimated free energies due to repeated use of the same input coordinates.

A scheme to reduce cycle closure errors is used in FEP+

https://doi.org/10.33011/livecoms.2.1.18378

whereby calculated free energy changes along the nodes of the network are re-sampled assuming estimates of the calculated free energy change along a node may be obtained from a Gaussian distribution centered on the estimated free energy change and with a standard deviation equal to the estimated standard deviation of the free energy change. The procedure then uses a maximum likelihood method to find new sets of free energy changes that minimize cycle closure errors [22]. An alternative approach computes the free energy change between a target and reference compound as a weighted average over all unique paths in the network, with the weights derived from the propagated uncertainties of each node [9]. Approaches as illustrated by Yang et al. for perturbation map design can also be used to compute relative free energies between target and reference compounds [19].

## 6.2 Best practices for reporting data

FC: @JS/ OpenFE people to contribute?

FC: @JS / OpenFE team/ Volunteers to expand on utility/ limitations of common metrics. Also see @Michael Gilson's GitHub issue.

Following best practices for data generation and their analysis does not mean that data is reported in the optimal way. As a practitioner of alchemical free energy simulations you also should use best practices for reporting and plotting your results. We encourage the following standard set of analyses and ways to represent data.

### Statistics to include in reporting data

As with any modelling technique, misuse of statistical analysis can skew the perception of how well models perform in free energy predictions. First, error estimates should always be included on your predictions in whatever form you present your data (scatterplots, barplots, etc; see next paragraph). We recommend performing triplicates of your predictions at minimum, with starting points that are expected to be uncorrelated, to ensure some measure of reliability in your data. This replication may seem excessive, but uncertainty estimates often underestimate the true statistical uncertainty. Where performing multiple replicas of the simulation is not possible, an error estimate from e.g. MBAR can be used, though bearing in mind this is likely an underestimated error.

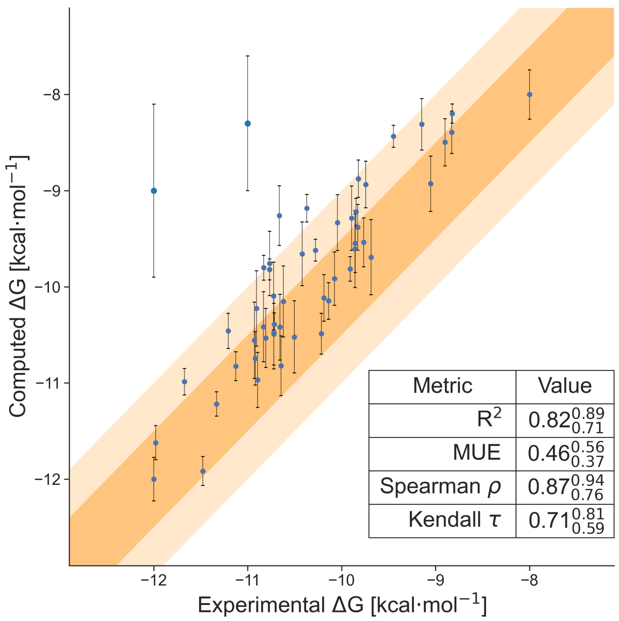
As alchemical free energy methods are used in drug discovery to quantify and rationalise structure activity relationships (SAR), the models ability to (a) correlate well with experiment and (b) rank-order the molecules by affinity, should both be computed. Conventionally, this means including an $R^2$ (or Pearson's R), where $R = +1$ means high correlation, $R = 0$ means no correlation, and $R = -1$ means high anti-correlation) and a Kendall τ (with perfect ranking agreement when τ=1 and perfect disagreement when τ=-1) metric in your results. Additionally, practitioners may choose to include a Spearman ρ as well. Brown et al. [33] have provided a useful analysis in terms of upper bounds of expected possible correlations between experiment and computation with a given potency range for the compounds. For example, for potency ranges of 2 log units it would be impossible to get a higher correlation in R than 0.8 because of experimental uncertainties [33]. What often is neglected to include is an error analysis on correlation statistics that arise from the errors of both experimental and computed data. One way to include such error analysis for correlation metrics is using bootstrapping on the datasets. The D3R community challenges follows best practices on their data evaluation with readily available python scripts online [34], based on work by Pat Walters [35]. Other analysis software also provide similar functionality for bootstrapping datasets [36].

Mean unsigned error (MUE, also called mean absolute error/MAE) is another key statistic to include in your results. Even though some models' near-perfect correlation and ranking statistics might suggest excellent accuracy, MUE values can still have errors of multiple kcal/mol, providing important additional insight into performance. Furthermore, MUE allows for unbiased comparisons between predictive models as it is less sensitive to dataset size. Other metrics such as Gaussian Random Affinity Model (GRAM) [37], Predictive Interval (PI) and Relative Absolute Error (RAE), attempt to correct for the inherent potency range of a dataset, which can aid in comparing success between different targets. We recommend further reading on evaluation of computational models [33, 35, 38, 39].

Reporting the results of relative free energy calculations requires care. As shown in Fig. 1, relative free energies can be performed arbitrarily as a forward or a reverse process, and thus relative free energies may be reported as either positively or negatively valued. The consequence of the two possible signs for relative free energies is that correlation statistics (such as Pearson's R and Kendall τ) can be skewed depending on which sign is analysed. The issue of this inconsistency can be circumvented by either plotting all datapoints within a consistent quadrant [8], or by avoiding the use of correlation statistics for assessment of relative free energy calculations and instead measuring accuracy using RMSE and MUE, which are unaffected by choice of sign.

### Presenting your data

As essentially all alchemical free energy prediction schemes are regression problems, the preferred type of plot is a scat-



| Metric | Value |
|---|---|
| $R^2$ | $0.82^{0.89}_{0.71}$ |
| MUE | $0.46^{0.56}_{0.37}$ |
| Spearman $\rho$ | $0.87^{0.94}_{0.76}$ |
| Kendall $\tau$ | $0.71^{0.81}_{0.59}$ |

**Figure 2. An example of recommended practices for graphing alchemical free energy predictions.** This figure shows the relation between predicted and experimentally-determined Gibbs free energy in kcal/mol with standard errors as error bars. The dark and light-orange regions depict the 1- and 2-kcal/mol confidence bounds. Statistical metrics for the data are reported, with 95% confidence intervals determined by bootstrapping analysis. Extra care should be taken when investigating potential outliers further.

ter plot (see Fig. 2). Most alchemical free energy projects will look at 10-50 ligands. Any study with <10 ligands is more suitable for bar plots (with inclusion of error bars), and is unlikely to provide meaningful statistics. Any study with >50 ligands typically contains multiple protein targets to which alchemical free energies may perform better on some targets than others. Because of this, it is bad practice to place multiple datasets on the same plot as this can suggest high model accuracy even though the individual models perform less well [39].

As we are interested mainly in the linear relationship between the alchemical free energy predictions and the experimentally-determined affinity values, plots should be depicted with the same range on both axes (i.e. $x = y$) with a 1:1 aspect ratio, with units for both experiment and simulation converted to be the same. If this skews the plot to a point where it is difficult to read of information, using the same dimensions, such that e.g. 1 cm is 1 kcal/mol is acceptable. Furthermore, bounds should be depicted for the 1- and 2-kcal/mol confidence regions. These regions can serve as tools to communicate your model performance: any predictions inside the 1 kcal/mol region can be seen as highly reliable, any predictions inside the 2 kcal/mol region should be seen as somewhat reliable, and any predictions outside the confidence regions should be expected to be unreliable and handled as outliers. In a drug discovery context, this type of data depiction may suggest the reliability of alchemical FE predictions in the project, and can give an idea of how trustworthy predictions can be for synthesis ideas. It is also recommended to included experimental error bars in all plots.

An example of a best practice scatter comparison between computed and experimental values is shown in Fig. 2, highlighting outliers, error bars and confidence intervals. The data for this plot is artificially generated for illustration purposes.

## 7 Alchemical free energy datasets: an overview

The following contains a non-exhaustive summary of alchemical free energy datasets that can serve as a starting point to review approaches or test new implementations. The field is moving towards a more standardised way of generating protein-ligand benchmark datasets and the progress of these efforts can be tracked here: https://github.com/openforcefield/FE-Benchmarks-Best-Practices. Currently lacking an exhaustive set of benchmark datasets, the review by Williams-Noonan et al. [40] contains an overview of recently published alchemical free energy studies. For comparison of FEP+ and Gromacs (using the AMBER99SB-ILDN and GAFF2 force field), cf. the recently published study by Pérez-Benito et al. [8]. An overview of further suggested benchmark sets can be found in the review by Mobley and Gilson [41] or on alchemistry.org [42]. These include cyclodextrins, the Cytochrome C peroxidase (CCP) protein model binding site, thrombin and bromodomains as well as solvation benchmark sets [43]. Please refer to table 1, for a small overview of datasets, what forcefields they used, and what the original study was it came from.

## 8 Checklist

FC: @Volunteer to update checklist once manuscript is nearing completion.

## KNOW WHAT YOU WANT TO SIMULATE

**Initial questions you should ask before you set up an alchemical free energy calculation using molecular dynamics simulations**

☐ Do I understand the biology, chemistry and physics of my system?
☐ Have I properly prepared my protein and ligand systems?
☐ Does my system contain any structures that require custom parameters?
☐ What simulation protocol will provide the most evidence to verify my hypothesis?
☐ Are the projected computational expense and runtime realistic for my scientific goals?
☐ Will my protocol be reproducible?
☐ Will my statistics be reliable? If not, would more replicates solve the problem?
☐ Can I open-source my data?

## PREPARING YOUR SIMULATIONS

**Steps to getting started setting up your alchemical free energy calculation**

☐ Make sure you know why you have picked your (combination of) force field(s)
☐ Energy minimize your system
☐ Equilibrate your system properly with your choice of thermodynamic ensemble
☐ Check the stability of your system and whether it behaves the way you believe it should

## RUNNING ABSOLUTE SIMULATIONS

**Steps to running your absolute alchemical free energy calculations**

☐ Check your ligands have the same, biologically correct binding pose
☐ Make sure your $\lambda$-scheduling is appropriate
☐ Check if your ligands are discharging and decoupling correctly
☐ Set up your restraints correctly
☐ Make sure you subsample the data in your free energy estimation protocol
☐ Apply the appropriate correction terms

## RUNNING RELATIVE SIMULATIONS

**Steps to running your relative alchemical free energy calculations**

☐ Check your ligands have the same, biologically correct binding pose
☐ Make sure your $\lambda$-scheduling is set correctly
☐ Make sure your molecular transformations are realistic (1-5 heavy atoms for reliable computations)
☐ Generate a perturbation network by your method of choice; check whether you have enough cycle closures to check consistency in the results
☐ Check whether dummy atoms were assigned correctly
☐ Consider subsampling the data in your free energy estimation protocol
☐ Apply the appropriate correction terms

https://doi.org/10.33011/livecoms.2.1.18378

## HOW DO I KNOW WHICH SIMULATIONS ARE UNRELIABLE?

**Situations suggesting your relative alchemical free energy calculations have not run properly (assuming absence of experimental affinities)**

☐ Standard error (σ) should not be >1 kcal·mol$^{-1}$
☐ Simulated systems have not converged - trajectories should be manually checked for consistency; other methods such as generating RMSD plots are also recommended

*Relative:*

☐ If you observe hysteresis in perturbations and incorrect cycle closures
☐ Energy differences >∼15 kcal·mol$^{-1}$ are likely unreliable

*Absolute:*

☐ Energies <∼-15 kcal·mol$^{-1}$ are likely unreliable
☐ The ligand has not sampled most of the intended region after the decoupling step
☐ The ligand is drifting out of the intended region after the decoupling step

## WHY ARE THEY NOT RELIABLE?

**Suggestions for finding out why your alchemical free energy calculations may not be reliable**

☐ Check again whether dummy atoms were assigned correctly
☐ Inspect the trajectories across the $\lambda$-schedule (particularly the endpoints) for problems described in the text
☐ Inspect the overlap matrices for lack of overlap

## DATA ANALYSIS

**Steps to analyzing your output data correctly**

☐ Make sure you have run enough replicates to ensure statistical reliability (>3)
☐ Compute both correlation and ranking coefficients and ranking statistics (e.g. r, ρ, MUE and τ)
☐ Include error bars in all your visual analyses

**Table 1.** Selection of example datasets

| Publication | Targets | Ligands | Force Field |
|---|---|---|---|
| D3R Grand Challenges [44] | | | |
| GC3 [45] | 6 | 266 | various |
| GC2 [46] | 1 | 102 | various |
| GC2015 [47] | 2 | 215 | various |
| SAMPL Challenges [48] | | | |
| SAMPL6 [49] | 3 | 21 | various |
| SAMPL5 [50] | 3 | 22 | various |
| SAMPL4 [51] | 2 | 23 | various |
| Schrödinger Datasets | | | |
| FEP+ Dataset [13] | 8 | 199 | OPLS2.1 |
| FEP+ Dataset [52] | 8 | 199 | OPLS3 |
| FEP+ Dataset [53] | 8 | 199 | OPLS3e |
| FEP+ Dataset [15] | 8 | 199 | GAFF 1.8 |
| FEP+ Dataset [16] | 8 | 199 | various |
| FEP+ Dataset [14] | 8 | 199 | GAFF2.1 |
| Fragments [54] | 8 | 96 | OPLS2.1 |
| Scaffold Hopping [28] | 6 | 21 | OPLS3 |
| Scaffold Hopping [14] | 6 | 21 | GAFF2.1 |
| Macrocycles [55] | 7 | 33 | OPLS3 |
| Further Suggested Datasets | | | |
| Cucurbit[7]uril (CB7) [41] | 1 | 15 | NA |
| Deep cavity cavitand [41] | 2 | 19 | NA |
| T4 Lysozyme [41] | 2 | 20 | NA |
| Merck set [56] | 5 | 169 | OPSL3 |

## Author Contributions

**ASJSM**: Coordinated the document, contributed to most sections, and co-designed Figs. **??**, **??**, **??**, 1, **??**, 2, and created Figs. **??**, **??**, **??**, **??** and replotted **??** and 1.
**BA**: Helped write the uncertainty estimation, stopping conditions, and output analysis sections and created figure **??**.
**HBM** Contributed to Sec. 6.2 and Fig. 2 and helped edit the paper.
**JDC**: Wrote Sec. **??** and **??** discussed structure and design of the whole document, suggested Figs. **??** and **??**.
**DFH** Contributed to Sec. **??** and helped edit the paper.
**MK**: Contributed to Sec. **??**, provided the data for figure **??**, compiled the dataset for Sec. 7 and helped edit the paper.
**JM**: Contributed to Sec. **??**, **??**, **??**, **??**, 6.1, and **??**
**DLM**: Contributed to the outline, drafted some of the sections, gave ideas on figures, and helped edit the paper.
**LNN**: Helped write the simulation length, stopping conditions, and information saving section. Edited and reviewed alchemical path section.
**SP** Wrote Sec. **??**.
**AR**: Created figure **??**, contributed to sections **??** and **??**, and helped edit the paper.
**JS**: Created Figs. **??**, **??**, **??**, **??**, **??**, 2, and an initial draft of 1. Wrote Sec. 6.2, the checklist Sec. 8, and contributed to general formatting discussions and editing.
**MRS**: Helped create figure **??**, wrote Sec. **??** describing choices for alchemical pathways and parts of **??** on the analysis for free energy calculations. Reviewed and edited text throughout.
**GT**: Contributed to Sec. 1 and 3, and helped edit the paper.
**HX**: Contributed Sec. 3.1, to Sec. **??**, and to Sec. **??**. For a more detailed description of author contributions, see the GitHub issue tracking and changelog at https://github.com/alchemistry/alchemical-best-practices.

## Other Contributions

Julia E. Rice participated in the original discussion of the document at the Best Practices in Molecular Simulation Workshop Hosted by at NIST, Gaithersburg, MD, August 24th-25th, 2017. Marieke Schor proofread the manuscript.

For a more detailed description of contributions from the community and others, see the GitHub issue tracking and changelog at https://github.com/alchemistry/alchemical-best-practices.

## Potentially Conflicting Interests

JM is a current member of the Scientific Advisory Board of Cresset. MK is employed by Cresset who commercially distribute a software for performing alchemical free energy calculations. MRS is a Open Science Fellow and consultant for Silicon Therapeutics. JDC is a current member of the Scientific Advisory Board of OpenEye Scientific Software and a consultant to Foresite Laboratories.

## Funding Information

ASJSM and JM acknowledge funding through an EPSRC flagship software grant: EP/P022138/1 MK and JM acknowledge funding through Innovate UK by KTP partnership 011120. AR acknowledges partial support from the Tri-Institutional Program in Computational Biology and Medicine and the Sloan Kettering Institute. HEBM acknowledges support from a Molecular Sciences Software Institute Investment Fellowship and Relay Therapeutics. DLM acknowledges support from the National Institutes of Health (R01GM108889, R01GM124270, and R01GM132386), and the National Science Foundation (CHE 1352608). JDC acknowledges support from the National Institutes of Health (NIH P30 CA008748, NIH R01 GM121505, R01 GM132386). A complete funding history for the Chodera lab can be found at http://choderalab.org/funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Author Information

**ORCID:**

Antonia S. J. S. Mey: 0000-0001-7512-5252
Bryce Allen: 0000-0002-0804-8127
Hannah E. Bruce Macdonald: 0000-0002-5562-6866
John D. Chodera: 0000-0003-0542-119X
Maximilian Kuhn: 0000-0002-2811-3934
Julien Michel: 0000-0003-0360-1760
David L. Mobley: 0000-0002-1083-5533
Levi N. Naden: 0000-0002-3692-5027
Samarjeet Prasad: 0000-0001-8320-6482
Andrea Rizzi: 0000-0001-7693-2013
Jenke Scheen: 0000-0001-9781-0445
Michael R. Shirts: 0000-0003-3249-1097
Gary Tresadern: 0000-0002-4801-1644
Huafeng Xu: 0000-0001-5447-0452
David F. Hahn: 0000-0003-2830-6880

## References

[1] **Mobley DL**, Klimovich PV. Perspective: Alchemical Free Energy Calculations for Drug Discovery. J Chem Phys. 2012; 137(23). doi: 10.1063/1.4769292.

[2] **Brown S**, Shirts M, Mobley D. Free Energy Calculations in Structure-Based Drug Design. In: ; 2010.p. 61–86.

[3] **Schindler CEM**, Baumann H, Blum A, Böse D, Buchstaller HP, Burgdorf L, Cappel D, Chekler E, Czodrowski P, Dorsch D, Eguida

MKI, Follows B, Fuchß T, Grädler U, Gunera J, Johnson T, Jorand Lebrun C, Karra S, Klein M, Knehans T, et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J Chem Inf Model. 2020; 60(11):5457–5474. doi: 10.1021/acs.jcim.0c00900.

[4] **Kuhn B**, Tichý M, Wang L, Robinson S, Martin RE, Kuglstatter A, Benz J, Giroud M, Schirmeister T, Abel R, Diederich F, Hert J. Prospective Evaluation of Free Energy Calculations for the Prioritization of Cathepsin L Inhibitors. J Med Chem. 2017; 60(6):2485–2497. doi: 10.1021/acs.jmedchem.6b01881.

[5] **Sherborne B**, Shanmugasundaram V, Cheng AC, Christ CD, DesJarlais RL, Duca JS, Lewis RA, Loughney DA, Manas ES, McGaughey GB, Peishoff CE, van Vlijmen H. Collaborating to Improve the Use of Free-Energy and Other Quantitative Methods in Drug Discovery. J Comput Aided Mol Des. 2016; 30(12):1139–1141. doi: 10.1007/s10822-016-9996-y.

[6] **Pérez-Benito L**, Keränen H, van Vlijmen H, Tresadern G. Predicting Binding Free Energies of PDE2 Inhibitors. The Difficulties of Protein Conformation. Sci Rep. 2018; 8(1):1–10. doi: 10.1038/s41598-018-23039-5.

[7] **Warren GL**, Do TD, Kelley BP, Nicholls A, Warren SD. Essential Considerations for Using Protein–Ligand Structures in Drug Discovery. Drug Discov. 2012; 17(23):1270–1281. doi: 10.1016/j.drudis.2012.06.011.

[8] **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. doi: 10.1021/acs.jctc.8b01290.

[9] **Mey ASJS**, Juárez-Jiménez J, Hennessy A, Michel J. Blinded Predictions of Binding Modes and Energies of HSP90-$\alpha$ Ligands for the 2015 D3R Grand Challenge. Bioorg Med Chem. 2016; 24(20):4890–4899. doi: 10.1016/j.bmc.2016.07.044.

[10] **Loeffler HH**, Michel J, Woods C. FESetup: Automating Setup for Alchemical Free Energy Simulations. J Chem Inf Model. 2015; 55(12):2485–2490. doi: 10.1021/acs.jcim.5b00368.

[11] **Hedges L**, Mey A, Laughton C, Gervasio F, Mulholland A, Woods C, Michel J. BioSimSpace: An Interoperable Python Framework for Biomolecular Simulation. J Open Source Softw. 2019; 4(43):1831. doi: 10.21105/joss.01831.

[12] **Gapsys V**, Michielssens S, Seeliger D, de Groot BL. Pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. J Comput Chem. 2015; 36(5):348–354. doi: 10.1002/jcc.23804.

[13] **Wang L**, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. J Am Chem Soc. 2015; 137(7):2695–2703. doi: 10.1021/ja512751q.

[14] **Kuhn M**, Firth-Clark S, Tosco P, Mey ASJS, Mackey M, Michel J. Assessment of Binding Affinity via Alchemical Free-Energy Calculations. J Chem Inf Model. 2020; 60(6):3120–3130. doi: 10.1021/acs.jcim.0c00165.

[15] **Song LF**, Lee TS, Zhu C, York DM, Merz KM. Using AMBER18 for Relative Free Energy Calculations. J Chem Inf Model. 2019; 59(7):3128–3135. doi: 10.1021/acs.jcim.9b00105.

[16] **Gapsys V**, Pérez-Benito L, Aldeghi M, Seeliger D, van Vlijmen H, Tresadern G, de Groot BL. Large Scale Relative Protein Ligand Binding Affinities Using Non-Equilibrium Alchemy. Chem Sci. 2020; doi: 10.1039/C9SC03754C.

[17] **Jespers W**, Esguerra M, Åqvist J, Gutiérrez-de-Terán H. QligFEP: An Automated Workflow for Small Molecule Free Energy Calculations in Q. J Cheminformatics. 2019; 11(1):26. doi: 10.1186/s13321-019-0348-5.

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

[19] **Yang Q**, Burchett W, Steeno GS, Liu S, Yang M, Mobley DL, Hou X. Optimal Designs for Pairwise Calculation: An Application to Free Energy Perturbation in Minimizing Prediction Variability. J Comput Chem. 2020; 41(3):247–257. doi: 10.1002/jcc.26095.

[20] **Xu H**. Optimal Measurement Network of Pairwise Differences. J Chem Inf Model. 2019; 59(11):4720–4728. doi: 10.1021/acs.jcim.9b00528.

[21] **Ciordia M**, Pérez-Benito L, Delgado F, Trabanco AA, Tresadern G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. J Chem Inf Model. 2016; 56(9):1856–1871. doi: 10.1021/acs.jcim.6b00220.

[22] **Wang L**, Deng Y, Knight JL, Wu Y, Kim B, Sherman W, Shelley JC, Lin T, Abel R. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. J Chem Theory Comput. 2013; 9(2):1282–1293. doi: 10.1021/ct300911a.

[23] **Robert A**, Lingle W, David LM, Richard AF. A Critical Review of Validation, Blind Testing, and Real- World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. Current Topics in Medicinal Chemistry. 2017; 17(23):2577–2585.

[24] **Cappel D**, Hall ML, Lenselink EB, Beuming T, Qi J, Bradner J, Sherman W. Relative Binding Free Energy Calculations Applied to Protein Homology Models. J Chem Inf Model. 2016; 56(12):2388–2400. doi: 10.1021/acs.jcim.6b00362.

[25] **Deflorian F**, Perez-Benito L, Lenselink EB, Congreve M, van Vlijmen HWT, Mason JS, de Graaf C, Tresadern G. Accurate Prediction of GPCR Ligand Binding Affinity with Free Energy Perturbation. J Chem Inf Model. 2020; doi: 10.1021/acs.jcim.0c00449.

[26] **Lenselink EB**, Louvel J, Forti AF, van Veldhoven JPD, de Vries H, Mulder-Krieger T, McRobb FM, Negri A, Goose J, Abel R, van Vlijmen HWT, Wang L, Harder E, Sherman W, IJzerman AP, Beuming T. Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. ACS Omega. 2016; 1(2):293–304. doi: 10.1021/acsomega.6b00086.

[27] **Chen W**, Deng Y, Russell E, Wu Y, Abel R, Wang L. Accurate Calculation of Relative Binding Free Energies between Ligands with Different Net Charges. J Chem Theory Comput. 2018; 14(12):6346–6358. doi: 10.1021/acs.jctc.8b00825.

[28] **Wang L**, Deng Y, Wu Y, Kim B, LeBard DN, Wandschneider D, Beachy M, Friesner RA, Abel R. Accurate Modeling of Scaffold Hopping Transformations in Drug Discovery. J Chem Theory Comput. 2017; 13(1):42–54. doi: 10.1021/acs.jctc.6b00991.

[29] **Hauser K**, Negron C, Albanese SK, Ray S, Steinbrecher T, Abel R, Chodera JD, Wang L. Predicting Resistance of Clinical Abl Mutations to Targeted Kinase Inhibitors Using Alchemical Free-Energy Calculations. Commun Biol. 2018; 1(1):70. doi: 10.1038/s42003-018-0075-x.

[30] **Albanese SK**, Chodera JD, Volkamer A, Keng S, Abel R, Wang L. Is Structure Based Drug Design Ready for Selectivity Optimization? bioRxiv. 2020; p. 2020.07.02.185132. doi: 10.1101/2020.07.02.185132.

[31] **Mondal S**, Tresadern G, Greenwood J, Kim B, Kaus J, Wirtala M, Steinbrecher T, Wang L, Masse C, Farid R, Abel R. A Free Energy Perturbation Approach to Estimate the Intrinsic Solubilities of Drug-like Small Molecules. . 2019; doi: 10.26434/chemrxiv.10263077.v1.

[32] **Liu S**, Wu Y, Lin T, Abel R, Redmann JP, Summa CM, Jaber VR, Lim NM, Mobley DL. Lead Optimization Mapper: Automating Free Energy Calculations for Lead Optimization. J Comput Aided Mol Des. 2013; 27(9):755–770. doi: 10.1007/s10822-013-9678-y.

[33] **Brown SP**, Muchmore SW, Hajduk PJ. Healthy Skepticism: Assessing Realistic Model Performance. Drug Discov. 2009; 14(7):420–427. doi: 10.1016/j.drudis.2009.01.012.

[34] Drugdata/Metk; 2018. Drug Design Data Resource.

[35] **Walters WP**. What Are Our Models Really Telling Us? A Practical Tutorial on Avoiding Common Mistakes When Building Predictive Models. In: *Chemoinformatics for Drug Discovery* John Wiley & Sons, Ltd; 2013.p. 1–31. doi: 10.1002/9781118742785.ch1.

[36] **Antonia M**, Michellab/Freenrgworkflows; 2019. michellab.

[37] **Cui G**, Graves AP, Manas ES. GRAM: A True Null Model for Relative Binding Affinity Predictions. J Chem Inf Model. 2020; 60(1):11–16. doi: 10.1021/acs.jcim.9b00939.

[38] **Jain AN**, Nicholls A. Recommendations for Evaluation of Computational Methods. J Comput Aided Mol Des. 2008; 22(3):133–139. doi: 10.1007/s10822-008-9196-5.

[39] **Walter P**, Some Thoughts on Evaluating Predictive Models;. http://practicalcheminformatics.blogspot.com/2019/02/some-thoughts-on-evaluating-predictive.html.

[40] **Williams-Noonan BJ**, Yuriev E, Chalmers DK. Free Energy Methods in Drug Design: Prospects of “Alchemical Perturbation” in Medicinal Chemistry. J Med Chem. 2018; 61(3):638–649. doi: 10.1021/acs.jmedchem.7b00681.

[41] **Mobley DL**, Gilson MK. Predicting Binding Free Energies: Frontiers and Benchmarks. Annu Rev Biophys. 2017; 46(1):531–558. doi: 10.1146/annurev-biophys-070816-033654.

[42] Alchemistry.org;. Accessed: 2019-08-01. http://www.alchemistry.org/wiki/Test_System_Repository.

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

[44] D3R Grand Challenges;. Accessed: 2019-08-01. https://drugdesigndata.org/about/grand-challenge.

[45] **Gaieb Z**, Parks CD, Chiu M, Yang H, Shao C, Walters WP, Lambert MH, Nevins N, Bembenek SD, Ameriks MK, Mirzadegan T, Burley SK, Amaro RE, Gilson MK. D3R Grand Challenge 3: Blind Prediction of Protein–Ligand Poses and Affinity Rankings. J Comput Aided Mol Des. 2019; 33(1):1–18. doi: 10.1007/s10822-018-0180-4.

[46] **Gaieb Z**, Liu S, Gathiaka S, Chiu M, Yang H, Shao C, Feher VA, Walters WP, Kuhn B, Rudolph MG, Burley SK, Gilson MK, Amaro RE. D3R Grand Challenge 2: Blind Prediction of Protein–Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. J Comput Aided Mol Des. 2018; 32(1):1–20. doi: 10.1007/s10822-017-0088-4.

[47] **Gathiaka S**, Liu S, Chiu M, Yang H, Stuckey JA, Kang YN, Delproposto J, Kubish G, Dunbar JB, Carlson HA, Burley SK, Walters WP, Amaro RE, Feher VA, Gilson MK. D3R Grand Challenge 2015: Evaluation of Protein–Ligand Pose and Affinity Predictions. J Comput Aided Mol Des. 2016; 30(9):651–668. doi: 10.1007/s10822-016-9946-8.

[48] SAMPL Challenges;. Accessed: 2019-08-01. https://samplchallenges.github.io/.

[49] **Rizzi A**, Murkli S, McNeill JN, Yao W, Sullivan M, Gilson MK, Chiu MW, Isaacs L, Gibb BC, Mobley DL, Chodera JD. Overview of the SAMPL6 Host–Guest Binding Affinity Prediction Challenge. J Comput Aided Mol Des. 2018; 32(10):937–963. doi: 10.1007/s10822-018-0170-6.

[50] **Yin J**, Henriksen NM, Slochower DR, Shirts MR, Chiu MW, Mobley DL, Gilson MK. Overview of the SAMPL5 Host–Guest Challenge: Are We Doing Better? J Comput Aided Mol Des. 2017; 31(1):1–19. doi: 10.1007/s10822-016-9974-4.

[51] **Muddana HS**, Fenley AT, Mobley DL, Gilson MK. The SAMPL4 Host–Guest Blind Prediction Challenge: An Overview. J Comput Aided Mol Des. 2014; 28(4):305–317. doi: 10.1007/s10822-014-9735-1.

[52] **Harder E**, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, Kaus JW, Cerutti DS, Krilov G, Jorgensen WL, Abel R, Friesner RA. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J Chem Theory Comput. 2016; 12(1):281–296. doi: 10.1021/acs.jctc.5b00864.

[53] **Roos K**, Wu C, Damm W, Reboul M, Stevenson JM, Lu C, Dahlgren MK, Mondal S, Chen W, Wang L, Abel R, Friesner RA, Harder ED. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. J Chem Theory Comput. 2019; 15(3):1863–1874. doi: 10.1021/acs.jctc.8b01026.

[54] **Steinbrecher TB**, Dahlgren M, Cappel D, Lin T, Wang L, Krilov G, Abel R, Friesner R, Sherman W. Accurate Binding Free Energy Predictions in Fragment Optimization. J Chem Inf Model. 2015; 55(11):2411–2420. doi: 10.1021/acs.jcim.5b00538.

[55] **Yu HS**, Deng Y, Wu Y, Sindhikara D, Rask AR, Kimura T, Abel R, Wang L. Accurate and Reliable Prediction of the Binding Affinities of Macrocycles to Their Protein Targets. J Chem Theory Comput. 2017; 13(12):6290–6300. doi: 10.1021/acs.jctc.7b00885.

[56] **MCompChem**, fep-benchmark; 2019. https://github.com/MCompChem/fep-benchmark, [Online; accessed 9. Dec. 2019].