

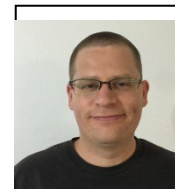
A critical review of validation, blind testing, and real-world use of alchemical protein-ligand binding free energy calculations

Robert Abel^{a*}, Lingle Wang^a, David L. Mobley^c and Rich A. Friesner^b

^a*Schrödinger, Inc., 120 West 45th Street, New York, NY 10036, United States*

^b*Department of Chemistry, Columbia University, 3000 Broadway, New York, NY, 10027, United States*

^c*Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine, CA, 92697, United States*



Abstract: Protein-ligand binding is among the most fundamental phenomena underlying all molecular biology, and a greater ability to more accurately and robustly predict the binding free energy of a small molecule ligand for its cognate protein is expected to have vast consequences for improving the efficiency of pharmaceutical drug discovery. We briefly review a number of scientific and technical advances that have enabled alchemical free energy calculations to recently emerge as a preferred approach, and critically consider proper validation and effective use of these techniques. In particular, we characterize a selection bias affect which may be important in prospective free energy calculations, and introduce a strategy to improve the accuracy of the free energy predictions.

Keywords: Computer-aided drug design, FEP, free energy, drug discovery, structure-based drug discovery, molecular dynamics, TI, thermodynamic integration, alchemical free energy calculations, protein-ligand binding

1. INTRODUCTION

Protein-ligand binding is of central importance in molecular biology, and directly mediates cellular metabolism, signal transduction, and coagulation among many other critical biological processes. Likewise, the vast majority of small molecule drug therapies achieve their desired affect through potent and selective binding to a relevant protein target. As such, there has been nearly a half-century of sustained interest in developing accurate, reliable, and precise methods to estimate ligand binding affinity.

Recently, physically rigorous free energy calculations have emerged as a preferred approach, and are beginning to see widespread adoption throughout the pharmaceutical industry.¹⁻⁴ The

progress toward this tipping point however has been gradual, and a wide variety of intellectual and technical contributions have been critical to the attainment of this long-standing goal. Looking back to this tremendously productive half-century of work, critical contributions have included

1. Formulation of physically rigorous estimators for the free energy change associated with an alchemical transformation that are amenable to being sampled via appropriate computer simulations;⁵⁻⁹
2. Development of molecular dynamics simulations techniques, and their extension to proteins;^{10,11}
3. Construction of highly accurate protein and small molecule molecular mechanics force fields;¹¹⁻²⁵

4. Derivation of sophisticated enhanced sampling methods, which may accelerate the sampling efficiency of bimolecular simulations while strictly satisfying detailed balance;²⁵⁻³⁰
5. Porting of molecular dynamics to GPU hardware, which provides in many applications up to 100x acceleration versus traditional CPU-based approaches.³¹⁻³³

Further, many of the practical and technical considerations associated with free energy calculations methods have recently been reviewed in detail in references 34 and 35. Here, however, we wish to turn our attention to what we consider to be equally important but often neglected subject: the validation and use of these methods, and how the nature of the use and goals may have a direct and sometimes counterintuitive effect upon the observed results.

2. FREE ENERGY CALCULATION VALIDATION AND USE

Pioneering work of the McCammon, Jorgensen, and Kollman groups from the late-80's to the early 90's brought together variety of advances to allow for the first protein-ligand binding free energy calculations to be performed.³⁶ These early contributions include the first calculation of a relative solvation free energy of two ligands by an alchemical transformation process,³⁷ the first relative binding free energy calculation of two ligands binding to a protein,³⁸ as well as the first prospective calculation of a relative binding affinity ahead of the experimental determination of the quantity.³⁹ However, due to the immense computational cost associated with these simulations, generally only one or a few calculations were reported in each publication, which made any real investigation of the accuracy of these methods impossible. Thus this early work, although filled with great promise, left an open question about what type of accuracy might actually be achievable with these methods if it were feasible to deploy them at a scale comparable to what would be needed to support an industrial drug discovery project.

Many free energy studies continue to focus on a small number of test cases, particularly when introducing new technology into the simulation protocols.⁴⁰⁻⁴² However, within the last several years, computer hardware advances have enabled multiple large-scale studies involving hundreds of ligands and multiple protein targets to be reported, which may enable now a real sense of the accuracy and reliability of these methods to be more fully understood.^{1,4,12,43,44} Furthermore, both blind testing, and detailed reports of actual use in discovery projects, are now commonly being reported, which should allow for a much greater understanding of the accuracy, reliability, and domain of applicability of these methods.^{1-3,45-51}

Thus, given this encouraging turn of events, one might now be able to delineate three basic categories of protein-ligand binding free energy calculation reporting manuscripts

1. Small-scale proof-of-concept articles with only a single or a few data point(s), which might be appropriate if the primary focus of the article is the reporting of a novel approach, and the reported calculations are intended to show that a computer implementation of the method is indeed possible, rather than accuracy;
2. Large-scale testing of a novel approach or enhancement of an existing approach on tens to hundreds of ligands across multiple target classes to profile the accuracy of the method within a particular domain of applicability;
3. Actual prospective use of free energy calculations to inform decisions to synthesize top scoring compounds at the exclusion of the synthesis of other lower scoring compounds.

Note, in this scheme, unblinded retrospective testing, blinded retrospective testing (such as that being enabled by the Drug Design Data Resource), as well as prospective testing are grouped together. Although the authors fully endorse the view that blinded testing (prospective or retrospective) is useful to ensure that one has not unintentionally designed the protocol to require information that might be unavailable when performing the

calculation in a more realistic context, blinded retrospective testing should generally look no different than unblinded retrospective testing so long as good methods development practices have been used. To be clear, if good methods' development practices are used, then unblinded and blinded testing results should be within the statistical error of each other unless one has been using the target data set for methods development rather than testing, which may have yielded over fitting. Given the vast data sets now available to most researchers through the PDB,⁵³ BindingDB,⁵⁴ ChEMBL,⁵⁵ D3R,⁵⁶ and other resources, we believe that any methods development group should have ready access to suitable test sets

However, in contrast, true prospective use of these methods raises data analysis challenges quite distinct from those posed by simple testing, blinded, prospective, or otherwise; and some of the challenges may be counterintuitive if one has not encountered them in other contexts. To be clear, by prospective use, we intend to describe a situation where N molecules are scored, and then a top scoring subset M , where $M \ll N$, is synthesized and assayed on the basis of the predicted affinities. We note here that a great deal of relevant learning can be taken from the virtual screening literature, where prospective applications are common.⁵⁶⁻⁵⁸ As detailed in the following section, the use of free energy calculations to compute a quantitatively accurate relative or absolute binding affinity raises unique challenges distinct from what is usually observed in virtual screening studies where the goal is typically simple classification; especially if one seeks to infer the numerical accuracy of the calculation methods from the prospective testing.

3. EVALUATION OF PROTEIN-LIGAND BINDING FREE ENERGY CALCULATIONS TO INFORM DECISION MAKING

Given the increasing frequency of reports of prospective free energy calculations guiding inhibitor design, we believe it is timely to consider how such studies should be evaluated.

This is very similar in purpose to important prior work by Brown et al. considering what quality of modeling results are feasible to obtain given the nontrivial noise that exists in most experimental binding affinity data,⁵⁹ as well as work by Mobley et al. detailing the methodological accuracy that might be needed to be impactful when working in a prospective setting.⁶⁰

To make the fundamental issues more transparent, we constructed a simple toy model system of 1000 compounds where;

1. The experimental pK_i values of the compounds are randomly drawn from Gaussian distribution with mean $pK_i = 6$ and $\text{stdev}(pK_i) = 1.5$
2. The predicted pK_i values of the compounds have exactly a 0.8 log unit root-mean-square error versus the experimental values, also following random assignment from an appropriate Gaussian distribution.

This toy model system, although very primitive, recapitulates many of the features of actual discovery projects. First, only ~10% of compounds in the set will have experimental affinities < 10 nM, which is generally needed for a molecule to be efficacious in a drug discovery setting; and second, the majority of molecules will have unremarkable experimental potencies ranging from double-digit micromolar to double-digit nanomolar, as would be expected to be observed in most drug discovery projects.⁶¹ Lastly, the accuracy of the predictions depicted here is comparable to those reported in most recent studies; and the number of idea molecules considered in the toy model is comparable to the number of idea molecules that might be considered by a project team utilizing commercially available reagents to explore synthesis possibilities at a given R-group attachment point.^{1,12,43} Excel files allowing for reconstruction of this data have been made available in supplemental information.

One can imagine this toy model system corresponding, for example, to the calculation of the absolute binding affinities of 1000 molecules generated by an R-group library

scan. We depict in figure 1.A the agreement of the toy model calculated data with the experimental data if all the molecules in the set were to be synthesized, and in figure 1.B the agreement of a subset flagged for synthesis on the basis of the computed potency values. Note, the points depicted in figure 1.B constitute simply the subset of points in figure 1.A that would be flagged for synthesis on the basis of the calculated potency values being found to less than 10 nM.

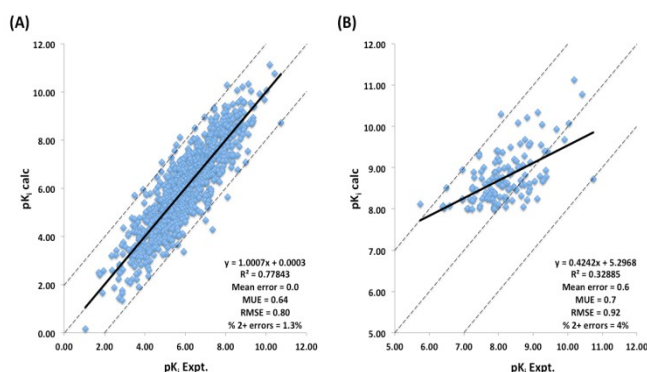


Figure 1. For a toy model system, in subpanel (A) we depict the agreement of the experimental affinities with the computed affinities for all the compounds in the set, and in subpanel (B) we depict the agreement for the subset of 128 compounds that would have been prioritized on the basis of the computed values.

Several features of these data are quite striking. Perhaps the most immediate observation is all routinely used accuracy measures are quite different between the full set and the prioritized subset, and are uniformly worse for the set prioritized for synthesis. The apparent R^2 value has dropped from a highly predictive value of 0.78 to a mediocre value of 0.33, the RMSE has increased by more than a tenth of a log unit, and perhaps most disconcertingly, the frequency of large outliers has increased by more than 300%. This suggests our selection of the best binding compounds for synthesis has somehow diminished the apparent accuracy of the calculation. Note, in a real computation-driven application, only the data represented in figure 1.B would ever be obtained in the experimental work, and grossly incorrect conclusions could easily be drawn regarding the actual accuracy of the prediction method.

One should further note that the effect becomes even more severe as the synthesis criteria (in particular the potency cutoff) become more stringent. For example, if only the top five scoring compounds were selected for synthesis the apparent RMSE would balloon to 1.37 log units, nearly double the actual value on the full set, and the frequency of >2 log unit outliers grows to 20%. Thus when considering the accuracy of a prediction method prospectively, extreme care will need to be taken to ensure the biased selection of compounds for synthesis is not influencing the characterization of the accuracy of the prediction method.

A simple but intuitive way to understand this effect is to consider the following limiting case. If by chance one has a single highly erroneous calculation in the set where a particular molecule with experimental pKi of ~6 might have been incorrectly predicted to have a pKi value of 11, the synthesis of the molecule is essentially guaranteed (at least in the absence of other considerations). But in contrast a great many accurately predicted molecules where both the experimental pKi and predicted pKi are both ~6 will be deprioritized for synthesis. Thus, the accuracy statistics of the set of molecules ultimately synthesized will be skewed toward those few molecules that have been severely overpredicted, with that bias growing the more stringent one is with the synthesis criteria.

A visualization of this selection bias effect for a single molecule drawn from our toy model is depicted in figure 2. In this figure the experimental potency distribution from which the molecule is drawn is depicted in black, and the uncertainty in the calculated value (ie, the free energy calculation method RMSE) is depicted in blue. Several features are immediately apparent. First, for molecules where the computed value is extreme with respect to the distribution of experimental values, the observed error will not be unbiased. For example, as depicted in figure 2, imagine a particular compound is computed to have a potency of pKi = 11 with a computational method having an RMSE of 1 log unit. In principle, if the free energy calculation method

is unbiased, then an over estimate of 1 log unit should be no more frequent than an underestimate of one log unit over a set of computed compounds. However, since molecules in the toy model with an experimental pKi value of 12 are so much more rare than molecules with an experimental pKi value of 10, the computed potency will in hindsight appear to overestimate the experimental potency considerably more often than it is underestimated.

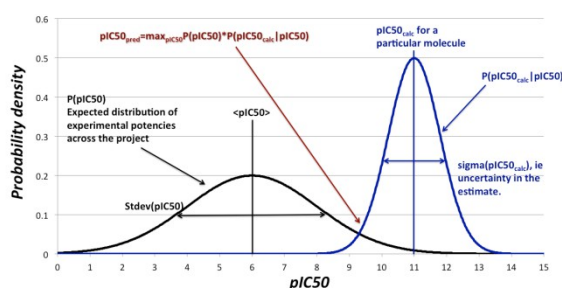


Figure 2. A graphical representation of the selection bias effect on the accuracy of the affinities is here depicted. The experimental potency distribution of the space of all compounds considered is depicted in black, the calculated potency and calculation error distribution is depicted in blue, and the Bayesian optimal predicted affinity of the compound is denoted in red.

Interestingly, visual analysis shown in figure 2 also suggests a straightforward way to correct for this selection bias effect. If the distribution of experimental potencies is somehow known, then it is a simple exercise in Bayesian probability to show the optimal predicted potency for the molecule is the value which optimizes the posterior distribution, i.e.

$$pIC50_{pred} = \max_{pIC50} P(pIC50) * P(pIC50_{calc} | pIC50) \quad (1)$$

where $pIC50_{pred}$ is the optimal prediction of the potency of the molecule, $pIC50_{calc}$ is the free energy calculation affinity value of the molecule, $P(pIC50)$ is the experimental affinity probability distribution of all the molecules in the set from which the particular predicted molecule was drawn, and $P(pIC50_{calc} | pIC50)$ is the conditional probability a molecule with a particular experimental $pIC50$ would result in a particular computed $pIC50_{calc}$ value.

Further, for our toy model system, this corrected predicted value ($pIC50_{pred}$), has a convenient closed form solution of

$$pIC50_{pred} = \frac{S^2_{\langle pIC50 \rangle} \cdot pIC50_{calc} + S^2_{pIC50_{calc}} \cdot \langle pIC50 \rangle}{S^2_{\langle pIC50 \rangle} + S^2_{pIC50_{calc}}} \pm \frac{S_{\langle pIC50 \rangle} \cdot S_{pIC50_{calc}}}{\sqrt{S^2_{\langle pIC50 \rangle} + S^2_{pIC50_{calc}}}} \quad (2)$$

To validate this simple correction to the calculated values, in figure 3 we have plotted the corrected predicted values using equation 2 versus the experimental data for all thousand compounds, as well as those predicted to bind more tightly than 10 nM after application of equation 2 in subpanel B. Interestingly, the mean error, MUE and RMSE are comparable both for the full set of compounds and the selected subset of compounds. Thus application of equation 2 has eliminated the selection of any particular subset from exhibiting error statistics deviant with the full set, irrespective of the particular way the predicted values are used to select the compounds for synthesis. The application of equation 2 also results in 56 fewer compounds being prioritized for synthesis using a 10 nM cutoff, versus the selection of compounds obtained with the uncorrected calculated values. One might also note, the R^2 value for the compounds shown in figure 3 subpanel B is worse than the R^2 value shown for figure 1 subpanel B. We expect the reason for this is simply the reduced experimental potency range of the compounds selected in figure 3 subpanel B, where fewer weakly binders have been selected for synthesis.⁵⁸

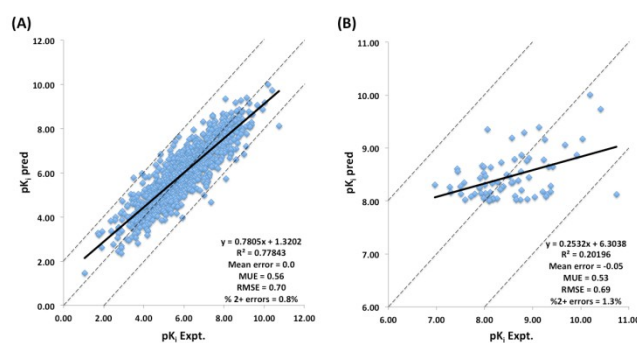


Figure 3. For a toy model system, in subpanel (A) we depict the agreement of the experimental affinities with the predicted affinities after application of the correction

described by equation 2 for all the compounds in the set, and in subpanel (B) we depict the agreement for the subset of 72 compounds that would have been prioritized on the basis of the predicted values yielded from application of equation 2.

However, application of equation 2 does pose some clear obstacles in more realistic prospective applications. Generally the true experimental distributions of the full set of compounds will not be known unless all the compounds are synthesized, which in turn would obviate the need to perform any scoring. However, the distribution could easily be estimated from synthesis of as few as 10 additional randomly selected compounds to provide an appropriate calibration. (Note, this is essentially the common statistical exercise of estimating the population variance from the sample variance.) An even more aggressive approach might be to assume the experimental potency distribution described by the toy model would be transferable from discovery project to discovery project.

As an initial attempt to investigate if such an *ad-hoc*, discovery-project-independent correction scheme might be useful in prospective applications, we present the prospective free energy calculation results from reference 51 in table 1. In that work, free energy calculations assisted with the identification of a compound with a binding affinity 1 full log-order tighter than any previously identified series compound. Further, synthesis of only 4 compounds prioritized by free energy calculations was required to identify this novel tight binding compound.

Interestingly, several features of the data presented in table 1 are consistent with our simple toy system. The mean error is large and positive indicating an apparent bias toward overestimating the binding affinities of the compounds, and the apparent RMSE and MUE are much larger than what was observed in retrospective free energy calculations for the same system. The toy model presented here suggests this apparent mismatch between the

retrospective and prospective accuracy may not be due to any unexpected methodological deficiency in the reported prospective free energy calculations, but rather instead due to an extreme selection bias effect manifested from the synthesis of only the top four scoring compounds in the considered set.

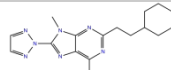
In table 1, we also report what predictions would have been made for the 4 prospectively prioritized compounds had equation 2 been used to correct the calculated values prior to the synthesis. Interestingly, because the authors of reference 51 report the cycle-closure estimate of calculation convergence (σ_{cc}),⁴² and the intrinsic accuracy of the force field used in that work (σ_{FF}) has been carefully profiled in earlier work to be ~ 0.9 kcal/mol in free energy calculation applications,¹² the σ_{calc} value appropriate to use in the application of equation 2 is necessarily compound specific, where

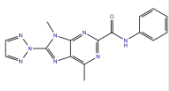
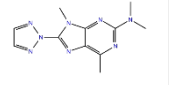
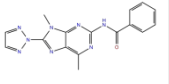
$$\sigma_{calc} = \text{sqrt}(\sigma_{cc}^2 + \sigma_{FF}^2) = \text{sqrt}(\sigma_{cc}^2 + 0.9^2) \quad (3)$$

and the σ_{cc} value will be different molecule to molecule.

Further, in order to apply equation 2 to these data, an assumption must be made regarding the experimental potency distribution of all compounds considered, including those explicitly not synthesized. A parsimonious assumption which can allow for direct application of equation 2 is to assume the activity distributions for this receptor may mirror the activity distributions generally seen in other discovery projects, ie $\langle \text{pKi} \rangle \approx 6$ and $\text{stdev}(\text{pKi}) \approx 1.5$.

Table 1. Prospective free energy calculation data for four GPCR inhibitors where the calculations were explicitly used to prioritize the synthesis of the four compounds.

Li ga nd	Structure	Δ G_E * xpt.	Δ G_{ca} * lc	σ_c *c	σ_c alc **	Δ G_{pr} ** ed	σ_{pr} ** ed
8		- 12.	- 11.	0. 8	1. 2	- 10.	1. 07

		16	90	8	6	90	
11		- 10. 40	- 12. 89	1. 5 2	1. 7 7	- 10. 91	1. 34
13		- 8.3 1	- 9.2 9	0. 4 7	1. 0 2	- 9.0 8	0. 91
17		- 9.1 2	- 12. 86	1. 3 3	1. 6 1	- 11. 10	1. 27
Mean error:			- 1.74			- 0.50	
MUE:			1.87			1.13	
RMSE:			2.30			1.26	
R2			0.22			0.63	

*The reported experimental affinity values ($\Delta G_{\text{Expt.}}$) and the prospectively computed free energy calculation results (ΔG_{calc}) and calculation convergence estimate are taken from reference 51. The units of the reported free energies is kcal/mol.

**The derived results free energy calculation predictions ΔG_{pred} and calculation uncertainty estimates σ_{calc} and σ_{pred} are obtained by application of equations 2 and 3 to the earlier reported data. The units of the reported free energies is kcal/mol.

Interestingly, even in the absence of any project-specific knowledge regarding the true experimental potency distribution of the considered compounds, application of equation 2 yields an enormous improvement in the agreement of the predicted potencies with the experimental data. The mean error, which measures the tendency of the bias of the prospective calculations to overestimate the experimental affinities, has been decreased in absolute value from an egregious -1.74 kcal/mol to a more satisfactory -0.5 kcal/mol. Likewise, the RMSE has been reduced from 2.3 kcal/mol to 1.26 kcal/mol, and the R^2 value has also been increased from 0.22 to 0.63. The

reason for the dramatic effect of the correction on the observed R^2 value is due to the data reported for the most unconverged calculations, i.e. ligands 11 and 17, being perhaps unsurprisingly the most erroneous, where in contrast the Bayesian estimate by its construction shifts the unconverged data points more toward the mean value of the experimental potency distribution.

A second data set that may be used to investigate whether or not application of equation 2 may provide utility in more realistic applications is reported in reference 62. In that reference, results for 138 prospective free energy calculations are reported across 7 different discovery projects. Since no convergence error estimates were reported, a calculation uncertainty of 0.8 log units was used uniformly. Likewise, consistent with our toy model, an experimental distribution of $\langle \text{pKi} \rangle \approx 6$ and $\text{stdev}(\text{pKi}) \approx 1.5$ was assumed. In figure 4 we plot the originally reported predictions, and the corrected predictions obtained by way of application of equation 2. Application of the correction term has improved the mean error, MUE, and RMSE by several tenths of a log unit; and the frequency of large outliers, which may be particularly disruptive to a discovery project, has been reduced by 66%.

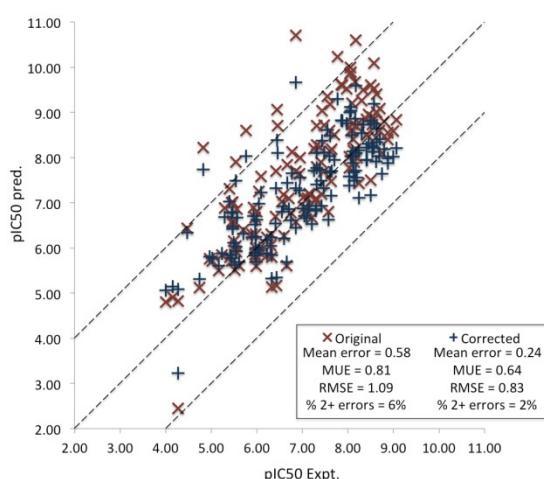


Figure 4. We here plot 138 agreement of 138 prospective FEP calculations used to prioritize synthesis versus the experimental data.⁶¹ The originally reported data is plotted in red, and the predictions obtained from application of equation 2 are plotted in blue. Error measures

are reported from both sets of data, however R^2 is omitted since the data has been collected from 7 different discovery projects.

The aforementioned notwithstanding, clearly this is insufficient prospective data from which to draw definitive conclusions regarding the effectiveness of equation 2 in practice, and whether or not the experimental potency distribution articulated in the construction of the toy model should be expected to approximately apply across many different discovery projects. However, we do believe the observation to be provocative, and look very forward to a great surge of prospective design work enabled by free energy calculations that might enable a more unambiguous consideration of these initial observations. In particular, when any computational technique is used to prioritize compounds for synthesis, a biased selection is being made, and this may have the unintended consequence of selection of compounds with large errors in their computed values in addition to good binders. Encouragingly, Bayesian approaches appear well suited to ameliorate this effect, even when project specific prior information is unavailable

Lastly, we wish to note that selection bias effects such as those detailed here may be relevant even if the project chemistry goals may be only to maintain the potency of tight binding matter while resolving some other ADMET liability. In such a situation, most new idea molecules designed to resolve the ADMET liability, e.g. adding a polar group to improve compound solubility, will likely diminish potency; and if only those few idea molecules designed to resolve the ADMET liability which are also predicted to maintain potency are synthesized, we expect a similar selection bias effect to what has been characterized above to be manifested in the resulting experimental data.

CONCLUSION

A variety of theoretical, technical, and practical advances are now positioning free energy calculations to become a standard approach to

inhibitor optimization and design both within academic laboratories and industrial drug discovery projects. Further, the maturity of the technology, availability of easy to use graphical interfaces, and the advent of low-cost high-performance GPU computing have placed the technology within reach of nonspecialists. As such, we expect within the next few years to observe a tremendous surge in the number of reported prospective applications, where free energy calculations are used to directly inform synthesis and design decisions.

In preparation for this expected body of prospective data, we have introduced a simple toy model that may help to elucidate some of the qualitative features we expect to later observe. Further, the toy model suggests there may be an opportunity to improve the accuracy of predictions from free energy calculation methods through more careful consideration of the expected experimental affinity distributions of all the compounds considered, including those explicitly deprioritized for synthesis on the basis of the free energy calculations. Although in no way definitive, the reanalysis of previously reported data does suggest these types of selection bias artifacts should indeed be expected to occur in real-world prospective applications, and a correction scheme of the type facilitated by equation 2 may be useful to improve prediction accuracy in realistic settings.

Further, we believe the correction scheme described here should have utility beyond simply the modeling of potency, and may be relevant any time one is using predictive calculations to optimize a property where idea molecules manifesting the desired extreme value of the property should be expected to be rare. As such, we expect generalizations of the approach described here in to address questions of ligand binding selectivity, clearance, and other such ADMET properties to be a fruitful future direction.

CONFLICT OF INTEREST

The authors declare the following competing financial interest(s): D.L.M. is a consultant to OpenEye Scientific Software. and serves on its Scientific Advisory Board. R.A.F. has a

significant financial stake in, is a consultant for, and is on the Scientific Advisory Board of Schrodinger, Inc.

ACKNOWLEDGEMENTS

We thank John Chodera and William Jorgensen for many helpful discussions.

SUPPLEMENTARY MATERIAL

An excel file detailing the construction of the toy model is included as supplementary information.

REFERENCES

- [1] L Wang, Y Wu, Y Deng, B Kim, L Pierce, G Krilov, D Lupyan, S Robinson, MK Dahlgren, J Greenwood, DL Romero, C Masse, JL Knight, T Steinbrecher, T Beuming, W Damm, E Harder, W Sherman, Mark Brewer, R Wester, M Murcko, L Frye, R Farid, T Lin, DL Mobley, WL Jorgensen, BJ Berne, RA Friesner, R Abel. 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. *Journal of the American Chemical Society*. **2015**,137(7):2695-703
- [2] Rombouts FJ, Tresadern G, Buijnsters P, Langlois X, Tovar F, Steinbrecher TB, Vanhoof G, Somers M, Andrés JJ, Trabanco AA. Pyrido [4, 3-e][1, 2, 4] triazolo [4, 3-a] pyrazines as Selective, Brain Penetrant Phosphodiesterase 2 (PDE2) Inhibitors. *ACS medicinal chemistry letters*. **2015**,6(3):282-6.
- [3] Lovering F, Aevazelis C, Chang J, Dehnhardt C, Fitz L, Han S, Janz K, Lee J, Kaila N, McDonald J, Moore W. Imidazotriazines: Spleen Tyrosine Kinase (Syk) Inhibitors Identified by Free-Energy Perturbation (FEP). *ChemMedChem*. **2016**,11(2):217-33.
- [4] Christ CD, Fox T. Accuracy assessment and automation of free energy calculations for drug design. *Journal of chemical information and modeling*. **2013**,54(1):108-20.
- [5] Kirkwood JG. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*. **1935**,3(5):300-13.
- [6] Zwanzig RW. High-temperature equation of state by a perturbation method. I. nonpolar gases. *The Journal of Chemical Physics*. **1954**,22(8):1420-6.
- [7] Bennett CH. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*. **1976**,22(2):245-68.
- [8] Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*. **2008**,129(12):124105.
- [9] Jarzynski C. Comparison of far-from-equilibrium work relations. *Comptes Rendus Physique*. **2007**,8(5):495-506.
- [10] Levitt M, Warshel A. Computer simulation of protein folding. *Nature*. **1975**,253(5494):694-8.
- [11] McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. *Nature*. **1977**,267(5612):585-90.
- [12] E Harder, W Damm, J Maple, C Wu, M Reboul, JY Xiang, L Wang, D Lupyan, MK Dahlgren, JL Knight, JW Kaus, D Cerutti, G Krilov, WL Jorgensen, R Abel, RA Friesner. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *Journal of Chemical Theory and Computation*. **2015**,12(1):281-96.
- [13] Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*. **2001**,105(28):6474-87.
- [14] Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved peptide and protein torsional energetics with the OPLS-AA force field. *Journal of chemical theory and computation*. **2015**,11(7):3499-509.
- [15] Jorgensen WL, Tirado-Rives J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of*

- the United States of America. **2005**,102(19):6665-70.
- [16] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*. **1995**,117(19):5179-97.
- [17] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*. **2006**,65(3):712-25.
- [18] Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*. **2010**,78(8):1950-8.
- [19] Cerutti DS, Rice JE, Swope WC, Case DA. Derivation of fixed partial charges for amino acids accommodating a specific water model and implicit polarization. *The Journal of Physical Chemistry B*. **2013**,117(8):2328-38.
- [20] Li DW, Brüschweiler R. NMR-based protein potentials. *Angewandte Chemie International Edition*. **2010**,49(38):6778-80.
- [21] MacKerell Jr AD, Bashford D, Bellott ML, Dunbrack Jr RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D. All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *The journal of physical chemistry B*. **1998** 102(18):3586-616.
- [22] Best RB, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, MacKerell Jr AD. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ_1 and χ_2 dihedral angles. *Journal of chemical theory and computation*. **2012**,8(9):3257-73.
- [23] Schuler LD, Daura X, Van Gunsteren WF. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry*. **2001**,22(11):1205-18.
- [24] Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry*. **2004**,25(13):1656-76.
- [25] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *Journal of computational chemistry*. **2004**,25(9):1157-74.
- [26] Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*. **1986**,57(21):2607.
- [27] Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*. **1999**,314(1):141-51.
- [28] Wang L, Friesner RA, Berne BJ. Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2). *The Journal of Physical Chemistry B*. **2011**,115(30):9431-8.
- [29] Wang L, Berne BJ, Friesner RA. On achieving high accuracy and reliability in the calculation of relative protein–ligand binding affinities. *Proceedings of the National Academy of Sciences*. **2012**,109(6):1937-42.
- [30] Kaus JW, Harder E, Lin T, Abel R, McCammon JA, Wang L. How to deal with multiple binding poses in alchemical relative protein–ligand binding free energy calculations. *Journal of chemical theory and computation*. **2015**,11(6):2670-9.
- [31] Michael Bergdorf, Sean Baxter, Charles A. Rendleman, and David E. Shaw. *Desmond/GPU Performance as of October 2015*. Technical Report, New York: D. E. Shaw Research Technical Report DESRES/TR, 2015.

- [32] Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of chemical theory and computation*. **2013**,9(9):3878-88.
- [33] Eastman P, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, Beauchamp KA, Lane TJ, Wang LP, Shukla D, Tye T. OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *Journal of chemical theory and computation*. **2012**,9(1):461-9.
- [34] Pohorille A, Jarzynski C, Chipot C. Good practices in free-energy calculations. *The Journal of Physical Chemistry B*. **2010**,114(32):10235-53.
- [35] Hansen N, Van Gunsteren WF. Practical aspects of free-energy calculations: A review. *Journal of chemical theory and computation*. **2014**,10(7):2632-47.
- [36] Kollman P. Free energy calculations: applications to chemical and biochemical phenomena. *Chemical reviews*. **1993**,93(7):2395-417.
- [37] Jorgensen WL, Ravimohan C. Monte Carlo simulation of differences in free energies of hydration. *The Journal of chemical physics*. **1985**,83(6):3050-4.
- [38] Wong CF, McCammon JA. Dynamics and design of enzymes and inhibitors. *Journal of the American Chemical Society*. **1986**,108(13):3830-2.
- [39] Merz Jr KM, Kollman PA. Free energy perturbation simulations of the inhibition of thermolysin: prediction of the free energy of binding of a new inhibitor. *Journal of the American Chemical Society*. **1989**,111(15):5649-58.
- [40] Robertson MJ, Tirado-Rives J, Jorgensen WL. Performance of Protein-Ligand Force Fields for the Flavodoxin-Flavin Mononucleotide System. *The Journal of Physical Chemistry Letters*. **2016** In press.
- [41] Gumbart JC, Roux B, Chipot C. Efficient determination of protein-protein standard binding free energies from first principles. *Journal of chemical theory and computation*. **2013**,9(8):3789-98.
- [42] 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. *Journal of chemical theory and computation*. **2013**,9(2):1282-93.
- [43] 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. *Journal of chemical information and modeling*. **2015**,55(11):2411-20.
- [44] Mikulskis P, Genheden S, Ryde U. A Large-scale test of free-energy simulation estimates of protein-ligand binding affinities. *Journal of chemical information and modeling*. **2014**,54(10):2794-806.
- [45] Mey AS, Juárez-Jiménez J, Hennessy A, Michel J. Blinded predictions of binding modes and energies of HSP90- α ligands for the **2015** D3R Grand Challenge. *Bioorganic & Medicinal Chemistry*. 2016, In press.
- [46] Rocklin GJ, Boyce SE, Fischer M, Fish I, Mobley DL, Shoichet BK, Dill KA. Blind prediction of charged ligand binding affinities in a model binding site. *Journal of molecular biology*. **2013**,425(22):4569-83.
- [47] Jorgensen WL. Efficient drug lead discovery and optimization. *Accounts of chemical research*. **2009**,42(6):724-33.
- [48] Jorgensen WL. Computer-aided discovery of anti-HIV agents. *Bioorganic & Medicinal Chemistry*. **2016**, In press.
- [49] Erion MD, Dang Q, Reddy MR, Kasibhatla SR, Huang J, Lipscomb WN, van Poelje PD. Structure-guided design of AMP mimics that inhibit fructose-1, 6-bisphosphatase with high affinity and specificity. *Journal of the*

- American Chemical Society. **2007**,129(50):15480-90.
- [50] MR, Erion MD. Calculation of relative binding free energy differences for fructose 1, 6-bisphosphatase inhibitors using the thermodynamic cycle perturbation approach. *Journal of the American Chemical Society*. **2001**,123(26):6246-52.
- [51] Lenselink, Eelke, Louvel, Julien, Forti, Anna, van Veldhoven, Jacobus, de Vries, Henk, Mulder-Krieger, Thea, McRobb, Fiona, Negri, Ana, Goose, Joseph, Abel, Robert, Van Vlijmen, Herman, Wang, Lingle, Harder, Edward, Sherman, Woody, IJzerman, Adriaan, Beuming, Thijs. *ACS Omega*, **2016**: Submitted.
- [52] Berman HM. The protein data bank: a historical perspective. *Acta Crystallographica Section A: Foundations of Crystallography*. **2008**,64(1):88-95.
- [53] Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*. **2007**,35(suppl 1):D198-201.
- [54] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*. **2012**,40(D1):D1100-7.
- [55] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*. **2004**,3(11):935-49.
- [56] R. Amaro, V. Feher, M.K. Gilson, S.K. Burley, *Drug Design Data Resource: An Open Resource to Advance Computer - Aided Drug Design*, (n.d.). <https://drugdesigndata.org> (accessed ?, 2016).”
- [57] Walters WP, Stahl MT, Murcko MA. Virtual screening—an overview. *Drug Discovery Today*. **1998**,3(4):160-78.
- [58] Schneider G. Virtual screening: an endless staircase?. *Nature Reviews Drug Discovery*. **2010**,9(4):273-6.
- [59] Brown SP, Muchmore SW, Hajduk PJ. Healthy skepticism: assessing realistic model performance. *Drug Discovery Today*. **2009**,14(7):420-7.
- [60] Shirts MR, Mobley DL, Brown SP. Free energy calculations in structure-based drug design. *Drug Design: Structure- and Ligand-Based Approaches*. **2010**:61-86.
- [61] Hann MM. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm*. **2011**,2(5):349-55.
- [62] Abel, Robert. Schrödinger Inc. **2014**. <https://www.schrodinger.com/newsletter/s/fep-progress-report-free-energy-calculations-drug-discovery>.

Received: March 20, 2014
20, 2014

Revised: April 16, 2014

Accepted: April