Title
Towards an automated computational pipeline to guide drug lead optimization

Author
Liu, Shuai

Publication Date
2015

UNIVERSITY OF CALIFORNIA,
IRVINE

Towards an automated computational pipeline to guide drug lead optimization

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Pharmacological Sciences

by

Shuai Liu

Dissertation Committee:
Professor David Mobley, Chair
Professor Ray Luo
Professor Douglas Tobias
Professor Weian Zhao

2015

DEDICATION

This work is dedicated to my grandmother Baolan Zhang

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge assistance and financial support for the work presented here on a by-chapter basis:

1. I thank the Louisiana Board of Regents Research Competitiveness and Research Enhancement Subprograms as well as the Louisiana Optical Network Initiative (supported by the Louisiana Board of Regents Post-Katrina Support Fund Initiative grant LEQSF(2007-12)- ENH-PKSFI-PRS-01), and the National Science Foundation under NSF EPSCoR Cooperative Agreement No. EPS-1003897 with additional support from the Louisiana Board of Regents.

2. I thank the National Institutes of Health (1R15GM096257-01A1) as well as the computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513. I also grateful to OpenEye Scientific Software for support for SAMPL, including for the meeting and for logistical help with the website, and in particular we would like to thank Matt Geballe for help with the website and submissions and for helpful discussions, and Paul Hawkins, Greg Warren, and Geoff Skillman for helpful discussions and pointers on analysis. We are also thankful to Tom Peat (CSIRO) and colleagues for the experimental data which made the integrase portion of SAMPL possible, and helped initiate SAMPL4.

3. I am grateful to Michael Shirts (University of Virginia) and Pavel Klimovich (UCI) for helpful discussions, and to John Chodera (Memorial Sloan Kettering Cancer Center) for both useful input and an initial MCSS tool that helped motivate the work. I acknowledge the financial support of the National Institutes of Health (1R15GM096257-01A1), the National Science Foundation LA-SiGMA program (EPS-1003897), the Louisiana Board of Regents Research Competitiveness and Research Enhancement Subprograms, and Children's Hospital of New Orleans.

4. I thank the National Institutes of Health (1R15GM096257-01A1, 1R01GM108889-01), and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513.

5. I thank the National Institutes of Health (1R15GM096257-01A1, 1R01GM108889-01) and the National Science Foundation (CHE 1352608), and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513.

I thank Dr. David Mobley for his support, guidance, and encouragement. I acknowledge his patience and immense knowledge which tought me how to do research. Moreover, his personality has always been an excellent example of how kind a human being could be. I have many thanks for all the members in the Mobley lab for their support and invaluable discussions and a special thank to Pavel Klimovich who was of great help for all five years. I would also like to thank my father Jiaquan Liu, my mother Yuhua Gao, my girlfriend He Wang and all other friends who encouraged and supported me during all these years.

Chapter 1 is minimally modified reprint of the material as it appears in D. L. Mobley, S. Liu, D. Cerutti, W. C. Swope, and J. Rice, “Alchemical prediction of hydration free energies for SAMPL”, special issue, J. Comput. Aided Mol. Design 2012, 26(5):551-562
Chapter 2 is minimally modified reprint of the material as it appears in D. L. Mobley, S. Liu, N. M. Lim, K. L. Wymer, A. L. Perryman, S. Forli, N. Deng, J. Su, K. Branson, A. J. Olson, “Blind prediction of HIV integrase binding from the SAMPL4 challenge”, in J. Comput. Aided Mol. Design 2014, 28(4):327-345
Chapter 3 is minimally modified reprint of the material as it appears in S. Liu, Y. Wu, T. Lin, R. Abel, J. Redmann, C. M. Summa, V. R. Jaber, N. M. Lim, and D. L. Mobley, “Lead Optimization Mapper: Automating free energy calculations for lead optimization”, in J. Comput. Aided Mol. Design 2013, 27(9):755-770
Chapter 4 is minimally modified reprint of the material as it appears in S. Liu, L. Wang and D. L. Mobley, “Is ring breaking feasible in relative binding free energy calculations?” in J. Chem. Inf. Model. 2015, 55(4): 722-735

CURRICULUM VITAE

Shuai Liu

EDUCATION

Doctor of Philosophy in Pharmacological Sciences — **June, 2015**
University of California, Irvine — *Irvine, CA, USA*

Master of Science in Chemistry — **July, 2012**
University of New Orleans — *New Orleans, LA, USA*

Bachelor of Science in Pharmaceutical Science — **June, 2010**
Shandong University — *Jinan, SD, P. R. China*

RESEARCH EXPERIENCE

Graduate Research Assistant — **2010–2012**
David Mobley's lab at University of New Orleans — *New Orleans, LA, USA*

Graduate Research Assistant — **2012–2015**
David Mobley's lab at University of California, Irvine — *Irvine, CA, USA*

TEACHING EXPERIENCE

Teaching Assistant — **2009–2010**
University of New Orleans — *New Orleans, LA, USA*

INTERNSHIP

Intern — **July, 2012– September, 2012**
Schrödinger LLC — *New York, NY, USA*

COMFERENCES AND WORKSHOP PARTICIPATION

Free energy methods in drug design — **May, 2012**
Vertex Pharmaceuticals — *Cambridge, MA, USA*

Free energy methods in drug design — **May, 2014**
Vertex Pharmaceuticals — *Cambridge, MA, USA*

ABSTRACT OF THE DISSERTATION

Towards an automated computational pipeline to guide drug lead optimization

By

Shuai Liu

Doctor of Philosophy in Pharmacological Sciences

University of California, Irvine, 2015

Professor David Mobley, Chair

Molecular dynamics (MD) simulations are a promising tool to guide drug lead optimization. But because these tools are applied prospectively in drug discovery, blind tests provide a key opportunity to validate these for real-world applications. I participated in the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) 3 blind test (chapter 1) in which I used different force fields to calculate hydration free energies and SAMPL4 (chapter 2) in which I analyzed pose prediction results from different groups using different computational tools. The results from these blind tests improve our understanding of the accuracy of current methods, allowing us to improve these methods.

Large-scale applications of MD simulations to drug discovery have been few, partly because of the difficulty of planning and setting up the simulations. For example, alchemical relative free energy (RFE) calculations have relatively high accuracy in predicting differences in binding between drug lead compounds and new derivatives which are sought to improve binding potency. But setting up RFE calculations for large sets of compounds has required far too much manual intervention to be practical. I helped develop an algorithm, LOMAP (chapter3), to automatically plan and set up these calculations. Resulting applications indicated that it could successfully reduce the time of planning RFE calculations. But in this project, we assumed that relative free energy (RFE) calculations involving ring breaking will

introduce substantial error, and we tried to avoid these calculations as much as possible. Later, we quantitatively calculated what these errors would be to confirm this (chapter4).

Beside binding free energy calculations, MD simulations can also be used to predict solubilities. We used free energy calculations (chapter5) to calculate relative solubilities and compared the results with experiment and with results from more empirical chemical engineering methods. We found that our approach is more accurate, despite its straightforward nature.

Long-term, we are working towards developing an automated pipeline to help guide key aspects of drug lead optimization. My work helped with understanding the accuracy of current techniques, improving their automation, and providing a new technique for predicting physical properties like solubilities.

Chapter 1

Alchemical prediction of hydration free energies for SAMPL

1.1 Abstract

Hydration free energy calculations have become important tests of force fields. Alchemical free energy calculations based on molecular dynamics simulations provide a rigorous way to calculate these free energies for a particular force field, given sufficient sampling. Here, we report results of alchemical hydration free energy calculations for the set of small molecules comprising the 2011 Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenge. Our calculations are largely based on the Generalized Amber Force Field (GAFF) with several different charge models, and we achieved RMS errors in the 1.4-2.2 kcal/mol range depending on charge model, marginally higher than what we typically observed in previous studies [122, 130, 121, 93, 123]. The test set consists of ethane, biphenyl, and a dibenzyl dioxin, as well as a series of chlorinated derivatives of each. We found that, for this set, using high-quality partial charges from MP2/cc-PVTZ SCRF RESP fits provided marginally

improved agreement with experiment over using AM1-BCC partial charges as we have more typically done, in keeping with our recent findings [123]. Switching to OPLS Lennard-Jones parameters with AM1-BCC charges also improves agreement with experiment. We also find a number of chemical trends within each molecular series which we can explain, but there are also some surprises, including some that are captured by the calculations and some that are not.

1.2 Introduction

Hydration free energies based on molecular simulations have amassed numerous applications over the last decade [172, 105, 156, 32, 155, 69, 122, 130, 121, 120, 157, 93, 158, 123, 94]. This widespread interest is grounded in the fact that hydration free energies provide a rigorous test of force field accuracy, to the extent that all system degrees of freedom are adequately sampled. By extension, these calculations are believed to provide a proxy for the accuracy that can reasonably be expected in binding free energy calculations. Though these expectations may be well-founded, systematic tests of hydration free energies were typically limited to amino acid sidechain analogs until very recently (i.e. beginning around 2007 [122]), with few tests on diverse sets of small molecules. Perhaps as a consequence, hydration free energies, which yield just a single number potentially influenced by all of the force field parameters, have not been directly incorporated into force field development.

The relationship between hydration free energies and ligand-receptor binding free energies is complex, but not to be overlooked. Binding of a ligand to a receptor partially removes the ligand from solvent, and partially desolvates the receptor binding site. If, in the best case scenario, bound state interactions between the ligand, protein, and environment are well predicted, then the error in computed binding free energies will be dominated by any error in the hydration free energy. If errors in modeling hydration free energies are consistent within a chemical series, binding free energy calculations may still yield accurate relative free energy predictions. However, to predict binding of individual compounds, accurate absolute hydration free energies are necessary as well. In either of these general scenarios, the accuracy of hydration free energies, relative or absolute, sets a lower bound on the total error of predicted binding affinities (in the absence of fortuitous cancellation of error between hydration and binding calculations). Indeed, RMS errors in calculated binding free energies have proven comparable to or slightly larger than errors observed in hydration free energy calculations [33, 124, 14, 49, 23].

While hydration free energies are important, the experimental data is still somewhat limited: the largest test sets typically span 200-500 small molecules covering a relatively narrow range of chemical space [121, 158, 123, 100] for which the molecules have been studied extensively. However, pharmaceuticals often have a variety of functional groups which are only sparsely represented in these sets, fostering interest in new benchmarks. Several hydration free energy "prediction"[1] challenges have focused on more diverse sets of small molecules [130, 58, 52, 93]. Depending on the nature of the test set, performance can be substantially worse on the diverse, polyfunctional molecules in these test sets than on more typical test sets [130, 120, 93]. In some cases, these tests provide new insights – for example, the Generalized Amber Force Field (GAFF) [179, 178] does not properly model hypervalent sulfur compounds and shows systematic errors in alcohols which grow with the number of hydroxyl groups [120].

Here, we report our results from the latest SAMPL hydration free energy challenge, which in this case focuses on three molecules and their chlorinated derivatives. This data set is more focused than those of previous SAMPL challenges, providing an especially good opportunity to gain new insight. In these calculations, we applied our "standard" hydration free energy approach based on AM1-BCC partial charges [78, 79] with the GAFF small

[1] These are perhaps not true predictions, as the data is typically available in the literature, but often in relatively obscure places, requiring individual attention from a skilled practitioner to extract hydration free energies and evaluate the accuracy of the data.

molecule parameter set [179, 178], but we also took several other approaches. We tested an alternate set of partial charges from MP2/cc-PVTZ SCRF RESP [3] fits following the success of our recent work [123]. Alternatively, we tried a new set of QM-based charges calculated using explicit solvent simulations of each molecule to develop a solvent reaction field used in performing a fit of charges. We also took another approach for assigning GAFF or GAFF-like parameters, using OEAnte [36]. Further, we tested some newly developed aromatic carbon Lennard-Jones parameters. Finally, after seeing the correct hydration free energies for the molecules in the SAMPL challenge, we generated new estimates using OPLS [84] Lennard-Jones parameters for two of our submitted charge models, as well as a third charge model augmented with virtual sites to improve the molecular mechanics fit of MP2/cc-pvTZ SCRF calculations. These numerous approaches help to answer three fundamental questions about molecular mechanical hydration free energies. First, which charge model excels? Second, which van der Waals model excels? Third, given a specific quantum mechanical description of the electrostatics, does the quality of the molecular mechanics approximation to the quantum data matter?

1.3 Method

1.3.1 Overview

Our approach for hydration free energy calculations uses molecular dynamics simulations based on a classical atomistic force field to sample small molecules in water and in the gas phase, as well as at a variety of partially interacting intermediate "alchemical" states spanning between these end states. Using data collected from these simulations (specifically, potential energy differences evaluated between pairs of simulations) we are able to compute hydration free energies, in this case using the Multistate Bennett Acceptance Ratio (MBAR) approach for analysis [151]. This approach gives correct free energies for the particular parameter set given adequate simulation time and sufficient phase space overlap between neighboring alchemical states. This is the same approach in spirit as that in our previous work [122, 130, 121, 120, 93, 123], with minor and mostly insignificant changes relating to software versions.

Here, we briefly detail our general protocol for free energy calculations, then discuss setup differences between the different approaches considered, and conclude with some analysis details.

1.3.2 General "standard" protocol

We used a pre-release version of GROMACS 4.6[2] containing free energy modifications implemented by Michael Shirts (University of Virginia) to allow energy evaluations for MBAR to be done within the code rather than needing to store trajectory snapshots for later evaluation. This was used both for simulation setup and data analysis. The small molecule set was taken as provided by the SAMPL organizers, and then the OpenEye OEChem Python toolkit and Omega were used to generate likely solvated conformations for each molecule and assign AM1-BCC partial charges, unless otherwise noted. (As is relatively common with AM1-BCC charge calculations, the charges were not based on the wave function of a structurally optimized configuration.) These final conformations and charges were stored to mol2 files and then the AmberTools 1.4 version of Antechamber was used to assign GAFF atom types. AmberTools' `tleap` was used to convert these to AMBER prmtop and crd files, which were converted to GROMACS format using `acpype`. Small molecules were then set up in GROMACS and, for the solute-in-water case, solvated in TIP3P [83] water in a dodecahedral simulation box with at least 1.2 nm from the solute to the nearest box edge. The number of water molecules was approximately 510 for typical ethane systems, 875 for

[2] the FEP branch of git, commit 9e5241b9dfc8503f887d54640183bc8e397af261

typical biphenyl systems, and 890 for typical dioxin systems. AMBER combination rules (arithmetic average for σ and geometric for ϵ) were used.

Simulations were run using Langevin dynamics, as previously [122, 121], and a timestep of 2 fs. Van der Waals interactions were gradually switched off between 0.9 and 1.0 nm, and an analytical correction was applied to the energy and pressure [154]. PME was used for electrostatics, as previously, with a real-space cutoff of 1.2 nm. LINCS [67] was used to constrain bonds to hydrogen. Each system and λ value (where λ is a parameter ranging between 0 and 1, where 0 corresponds to the unmodified system, and 1 corresponds to the end-state of the transformation where the solute has no partial charges and no Lennard-Jones interactions) were independently minimized for up to 1000 steps of steepest-descents minimization, followed by 10 ps of constant volume equilibration, then 100 ps of constant pressure equilibration (for the solvated case). Following constant pressure equilibration, box sizes were adjusted at each λ value by an affine transformation to ensure each λ value had the correct volume for the target pressure. After this, we conducted an additional 5 ns of constant volume production simulation at each λ, discarding the first 100 ps as additional "equilibration", as previously [121].

The parameter λ controls the transformation between end states. In this version of GROMACS, we use two separate λ values, one controlling modification of partial charges (turning solute partial charges to zero) and the second controlling modification of Lennard-Jones interactions (turning solute LJ interactions to zero). Here, λ_{chg} = [0.0 0.25 0.5 0.75 1.0 1.00 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0] and λ_{LJ} = [0.0 0.00 0.0 0.00 0.0 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1.0] [122] . Hence, partial charges are first turned off and then LJ interactions are turned off separately. Hydration free energies are calculated by first computing the free energy of turning of solute-environment interactions in gas (vacuum) and then subtracting the free energy of turning these off in water.

The above constitutes our "standard" protocol. However, we also tested a number of deviations from this protocol which focused mainly in three areas: (a) alternate partial charge models; (b) alternate Lennard-Jones parameters; and (c) alternate software for assigning parameters.

1.3.3 MP2/SCRF charge model

In addition to AM1-BCC partial charges, we tried an alternate charge model (as suggested by our previous work [123]) based on MP2/cc-pVTZ calculations done in Gaussian03 [47] with a self-consistent reaction field (SCRF) continuum electrostatic model to represent solvent (particularly, the Integral Equation Formalism Polarizable Continuum Model (IEF-PCM) [16, 113] where the cavity is represented by a united atom topological model applied on atomic radii from the UFF force field (UA0)). Geometry optimization was done at the same level of theory. Charges were fitted using RESP [3] as implemented in Antechamber, taking Gaussian output files as input. Here, we refer to this charge model as "MP2/SCRF" for brevity.

We used this charge model with the same Lennard-Jones parameters as in the "standard" approach. To obtain hydration free energies with this charge model, we used the Lennard-Jones component of solvation (which is calculated without partial charges on the solute) from the "standard" protocol above and repeated the charging component of the calculations with the new charge model.

1.3.4 Solvent background charge QM

In addition to fitting charges to reproduce the electrostatics calculated by MP2 quantum calculations in a polarizable continuum solvent, we fitted charges to reproduce electrostatics of similar MP2-level quantum calculations in an explicit solvent reaction field generated by time-averaged densities of TIP4P-Ew [74] water around each target molecule in a variety of conformations. This approach, to be described in detail in a later publication, follows the conclusions of Karamertzanis et al. [86]: a non-polarizable model of a molecule should produce an electrostatic field which is precisely the average of the electrostatic fields produced by the unpolarized molecule in vacuum and by the polarized molecule in solvent. The derivation rests on a fundamental assumption that a change in the molecular dipole and the solvent reaction field which induced it are consistent, but it provides some rationalization of why the existing fixed-charge water models all have dipoles roughly halfway between the solution-phase dipole of 2.6-3.0D and the vacuum dipole of 1.85D [91]. Although the fitting method was complex and will require extensive scripting in order to be convenient, the result is simply an alternative set of charges for each molecule which enter into the same molecular mechanics energy function as existing force fields. These charges will be referred to as ”MP2/ExpSQ” for ”explicit solvent charge” denoting the reaction field potential employed in MP2 calculations.

1.3.5 Updated LJ Parameters for GAFF

Relatively little work has been done on optimizing GAFF parameters based on simulations of small molecules, so the recent suggestion of new aromatic carbon Lennard-Jones parameters for GAFF was intriguing [177]. Hence, for the molecules in the set containing those parameters (all but the ethane series) we repeated our hydration free energy calculations for the AM1-BCC (“standard”) charge model with these new parameters. We call this approach “Standard-newLJ”.

1.3.6 OEAnte Parameterization

We typically assign GAFF parameters using Antechamber. However, there is some interest in developing a more easily-extensible and open tool for atom typing and parameter assignment, which led Richard Dixon at Vertex Pharmaceuticals to release “OEAnte” [36], a Python tool based on the OpenEye libraries which handles atom typing and parameter assignment. Currently, it can either assign some rough guess parameters based on internal rules, or GAFF parameters after being passed a GAFF parameter file. We tested OEAnte in both modes, expecting that the rough guess parameters might give marginally worse results than standard GAFF, and expecting the GAFF mode to agree nearly perfectly with results from GAFF parameters assigned by Antechamber. Essentially, this last test amounts to validating parameter assignments by testing on hydration free energies. We tested these two sets of parameter assignments with AM1-BCC partial charges as in our standard approach. We call these “OEAnte-GAFF” and “OEAnte-Default”.

1.3.7 OPLS Results

After submitting our SAMPL predictions, discussions with Charles Kehoe at the University of California, San Francisco [89, 90] suggested that OPLS parameters might yield better results for this series than GAFF parameters. Motivated by this discussion, we replaced Lennard-Jones parameters for the few atom types in our study with OPLS Lennard-Jones parameters and modified the combination rules (geometric for OPLS). We maintained GAFF bonded parameters, expecting that hydration free energies would be relatively insensitive to these. Our goal was mainly to test whether the OPLS LJ parameters performed better here. Along those same lines, we maintained the AMBER 1-4 scaling parameters (0.5 for LJ and 0.8333 for Coulomb interactions) rather than switching to the OPLS values (0.5 for each) since we minted the AMBER bonded parameters. With this modification, we repeated both our standard and MP2/SCRF calculations with OPLS LJ parameters. We will call these methods “Standard-OPLS” and “MP2/OPLS”.

At a philosophical level, this swapping of parameters between AMBER and OPLS should perhaps be taken with a grain of salt. Computed hydration free energies are of course correct given the model that we use, but they may not be representative of what we would obtain from the true OPLS force field. Particularly, torsional parameters are typically derived after LJ parameters and partial charges are assigned, and depend on the charge model and LJ parameters. This coupling in some respects impedes the ability to use OPLS LJ parameters with AMBER bonded terms, and vise versa. That said, there has been considerable inter-pollination of the AMBER and OPLS force fields, as a reading of the AMBER parameter files will show, and occasionally, adopting OPLS LJ parameters straight into AMBER leads to marked improvements in computed observables, such as for alkynes [121].

1.3.8 Error analysis

Uncertainties in computed free energies were taken as the standard error in the mean from the MBAR estimator [151] and took into account the autocorrelation time of the potential energies being evaluated. Uncertainties in statistics such as the RMS error, R^2, Kendall τ and average errors were computed using a bootstrapping procedure, wherein new hydration free energy datasets were constructed from the original datasets by choosing random entries, with replacement. Uncertainties were estimated for these as 95% confidence intervals computed from looking at the variation of these statistics over a minimum of 1000 bootstrap trials (bootstrapping was repeated until confidence interval estimates converged to a tolerance of less than 1% of their values when averaged over 20 bootstrap trials).

1.4 Results

1.4.1 Accuracy of blind predictions

Blind predictions submitted for the SAMPL challenge, based on GAFF bonded and van der Waals parameters and three different charge models, are summarized in Figure 1.1 and statistics are shown in Table 1.1[3] (the table includes all methods used, not just those applied predictively). Previously, we reported hydration free energies with RMS errors relatively near 1 kcal/mol [122, 130, 123, 121], but test sets containing third-row elements or atypical organic functional groups have been much more challenging [120, 93]. Here, depending on the charge model, our submissions had RMS errors from 1.4 to 2.2 kcal/mol. The combination of narrow target ranges, particularly among compounds in the dioxin and biphenyl series, and larger errors made this set difficult in terms of rank ordering. Kendall τ values, measuring the fraction of compound pairs with correct rank orderings, ranged from 0.38±0.04 to 0.43±0.04 for these predictions. This is low, but nonzero, so some trends are correctly predicted. Overall, the RMS error is particularly poor given the range of free energies – as was noted at the meeting following the SAMPL challenge, predicting a constant hydration free energy equal to the average of all compounds (if that value were known) would have yielded an RMS error of 1.4 kcal/mol [51]. However, this would also have led to a Kendall τ value and correlation coefficients which were substantially lower (zero).

Overall, all charge models approaches did relatively well at predicting trends within the ethane derivatives, but appear to have predicted the wrong trend in the dioxin derivatives (the narrow range of the experimental HFEs, coupled with the large experimental uncertainties on many of these values, give some uncertainty as to what the experimental trend truly is). No charge model predicted a discernable trend in the biphenyl derivatives, even though the experimental HFEs in this group are more certain and their range is more broad than

[3] A full table of all computed hydration free energies is available in the Supporting Information (Table 2).

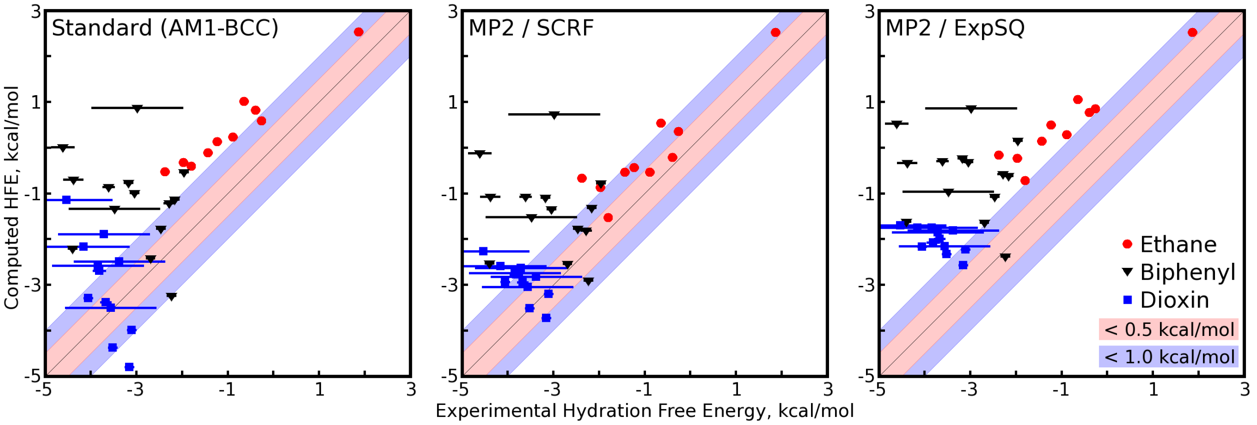


Figure 1.1: Computed versus experimental hydration free energies for several of the approaches and charge models tested here. Methods are discussed in the text and compounds are color coded by chemical series. Plots for additional approaches are shown in the Supporting Information. Inner gray bars denote 0.5 kcal/mol error, and outer gray bars denote 1.0 kcal/mol error.

that of the dioxins.

None of our SAMPL submissions is clearly superior. The MP2/SCRF charge fitting approach performed marginally better than our standard AM1-BCC charge assignment approach, as was expected (R^2 of 0.69 ± 0.04 kcal/mol and RMS error 1.59 ± 0.07 kcal/mol for the former versus 0.64 ± 0.04 kcal/mol and 1.88 ± 0.07 kcal/mol, respectively, for the latter). Kendall τ values were essentially within uncertainty of one another for all approaches. The MP2/ExpSQ charge fitting approach performed as well as the MP2/SCRF approach in predicting the overall trend in HFEs ($R^2 = 0.70 \pm 0.04$) but it consistently estimated the HFEs to be less favorable than in the experiment, leading to the highest overall RMS error (2.18 ± 0.07 kcal/mol). In fact, all of the SAMPL submissions, and also the other methods we tested after the competition, estimated hydration free energies with an unfavorable bias. This is consistent with our previous work (for example, across 504 molecules, the mean error was 0.67 kcal/mol in the same direction [121]) and clearly contributes to the RMS errors.

We also tested a new set of aromatic carbon Lennard-Jones parameters for GAFF [177]which had been informed by hydration free energy calculations. However, on this set (Standard-

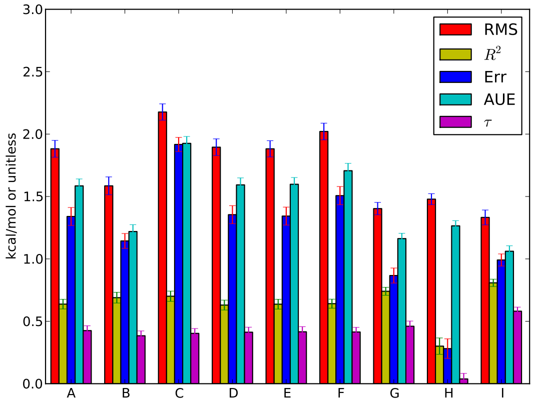


Figure 1.2: Statistics for the different approaches tested here. Shown are the RMS error, correlation coefficient R^2, average error ("Err"), average unsigned error (AUE), and Kendall τ (τ). Methods are (A) Standard; (B) MP2/SCRF; (C) MP2/ExpSQ; (D) Standard-newLJ; (E) OEAnte-Standard; (F) OEAnte-Default; (G) Standard-OPLS; (H) MP2/OPLS; (I) MP2/ExpSQ-Q; as discussed in the text. Error bars denote 95% confidence intervals. Data also in Table 1.1.

newLJ), these parameters did not result in a statistically significant change in the quality of calculated hydration free energies. Further, we applied the OEAnte package for assigning small molecule parameters as a substitute for Antechamber. The default set of parameters for this package are relatively crude and use far fewer atom types than GAFF (OEAnte-Default). The default parameter set in fact contained a misprint in the Lennard-Jones well depth for hydrogens which initially led to clearly incorrect hydration free energies. This was fixed and incorporated into the package prior to SAMPL. OEAnte can also be used to assign GAFF parameters, and we tested this approach as well (OEAnte-GAFF). When combined with AM1-BCC charges, the default approach resulted in marginally worse performance in terms of average error and RMS error (but comparable R^2 values) to our standard approach, while the GAFF parameter set produced results that were statistically indistinguishable from our standard approach, validating the package's parameter assignments.

1.4.2 Accuracy in follow-up testing after SAMPL

Several additional sets of hydration free energy calculations were performed after receiving the results of the SAMPL challenge. While many of our submissions to SAMPL tested different charge models, a significant focus in this follow-up testing was different Lennard-Jones parameters. As noted above, the Dill group found that OPLS Lennard-Jones parameters yielded better results than those with GAFF [90]. Combining these parameters with AM1-BCC partial charges improved the agreement with experiment (Standard-OPLS): RMS error of 1.4 ± 0.05 kcal/mol and an R^2 of 0.74 ± 0.03, though this is almost statistically indistinguishable from the MP2/SCRF approach. In contrast, combining OPLS Lennard-Jones parameters with charges assigned by MP2/SCRF (resulting in MP2/OPLS) decreased the RMS error slightly to 1.48 ± 0.04 kcal/mol but, by placing a favorable bias on the HFEs of the ethane derivatives in general and a severe favorable bias on the HFE of ethane itself, reduced R^2 to 0.30 ± 0.07. The Kendall τ value was also dramatically reduced in this case, from 0.46 ± 0.04 to 0.04 ± 0.04.

Finally, we looked at another possible source of error: the quality of the molecular mechanics approximation to electrostatic potentials derived from quantum-mechanical calculations. We repeated MP2/SCRF calculations on all molecules with chlorine substituents, saved the final checkpoint files containing the electron density, and then used the Gaussian cubegen utility to compute the electrostatic potential on rectangular grids around each molecule. The electrostatic potential created by a fitted charge model and the actual, target potential can then be directly compared, as shown in Figure 1.3. The RESP procedure is designed to minimize overall mean squared differences between the molecular mechanics potential and the quantum mechanical potential (subject to certain constraints) at a specified set of points, typically chosen at or near the molecular surface. As shown in Figure 1.3, the RESP-fitted charges may create localized regions of excess negative or excess positive potential, many of which are accessible to solvent molecules and particularly to solvent protons. By reviewing plots from chlorinated derivatives of ethanes, biphenyls, and dioxins, it became clear that the "knob" of excess negative potential and the “ring” of excess positive potential are typical features of RESP-fitted, atom-centered charges on organic chlorine. We therefore extended the charge model to include virtual sites attached to each chlorine atom, distal to the backbone carbon atoms and along the carbon-chlorine bond axis at a distance 30% of the carbon-chlorine bond length (roughly 0.54Å). Forces on the (massless) virtual sites can be computed during a typical molecular mechanics step and reassigned to atoms with mass by chain rules. We call this approach, with added charges at virtual sites, MP2/ExpSQ-Q.

By adding virtual sites to chlorine atoms and repeating the RESP fit, errors in the molecular mechanics approximation to the electrostatic potential were reduced by 40-50%, as shown in Table 3 of the Supporting Information. The virtual sites were pure charge sites with no Lennard-Jones interactions; in such cases it is always a concern that the unshielded charges may interact adversely with unshielded water protons. However, the charges on the virtual sites were always positive, reflecting the uniformity of the electrostatic potential defects around organic chlorine groups in atom-centered charge models, implying no risk of electrostatic singularities during the simulations. Overall, adding the extra points increased the negative charges on the chlorine atoms themselves, creating a stronger dipole along the carbon-chlorine bond than would have been fitted by an atom-centered charge model. While the fitted charges on chlorines did not always follow expected trends (see below) and the carbon-chlorine bonds were often modeled with small dipoles, these results can be understood as products of the RESP fit rather than the quantum-mechanical calculation.

The error in the molecular mechanics electrostatic approximation is readily quantifiable, but its effect on the computed hydration free energies is much less certain. We therefore performed additional free energy calculations (here with thermodynamic integration (TI)) to quantify the hydration free energy change implied by changing the atomic charges from the original atom-centered MP2/SCRF model to the new set augmented with virtual sites. A pre-release version of the mdgx program (part of the Amber software package [17]), augmented to run simulations involving virtual sites, was used to perform the TI. As shown in Table 2 of the supporting information, the $\Delta\Delta G$ values implied by changing the charges were generally modest, but the overall effect of adding the virtual sites was to reduce the RMS error to 1.33 ± 0.06 kcal/mol and improve R^2 to 0.81 ± 0.03 (correspondingly, the Kendall τ increased to 0.58 ± 0.03).

1.4.3 Key trends in hydration free energies

Overall, we expected that molecules in our series which are more highly chlorinated ought to be more favorably solvated, for two main reasons. First, adding a relatively electronegative chlorine tends to increase the overall polarity of the solute, and, in general, more polar molecules are more favorably solvated. Within the force field, this is typically represented by partial charges which become larger in magnitude (and typically negative on the chlorine atoms). Second, (for reasons which are related physically), chlorine atoms have stronger dispersion interactions with their environment than hydrogen atoms. Within the force field, this is seen by a substantially larger well depth for chlorine atoms as compared to hydrogen atoms ($\epsilon = 0.0150$ to 0.0157 kcal/mol for hydrogen, here in GAFF, versus $\epsilon = 0.2650$ for chlorine). These stronger dispersion interactions, we believed, would result in favorable *nonpolar* interactions with the surrounding solute that would more than offset the increased cavity volume due to the increased size of chlorine.

Both of these trends were generally borne out in our computed hydration free energies, though there are exceptions. Focusing on our MP2/SCRF results, and looking at a series of ethane derivatives, we find the expected trend generally holds up (Table 1.2). Adding chlorines does generally make solvation more favorable (Table 1.2(a)-(d)) in terms of both electrostatic and Lennard-Jones components of solvation. Because the effects contributing to this are largely local effects and strongly dependent on the first solvation shell, the solvent accessibility of the chlorines also makes a difference (Table 1.2(e)) – with an equal number of chlorines, the solute with the more solvent-accessible chlorines is preferentially solvated. In this specific case, there is also a dipole moment contribution – the overall dipole moment of the molecule is substantially higher in the case where the chlorines are at opposite ends, as well.

Dipole moment and charge distribution alone can make a substantial difference as well. Going from 1,1-dichlorethane to 1,1,1-trichloroethane results in a substantial overall decrease in dipole moment, and solvation becomes less favorable (Table 1.2(f)), both in our calculations and experimentally. Partial charges on the chlorine atoms are dramatically reduced, presumably because no nearby hydrogen atom donates electrons.

This isn't to say that the calculations correctly capture every trend. The 1,1-dichloroethane to 1,1,2,2,tetrachloroethane pair (Table 1.2(g)) is a mystery. Our computed charge sets (and the MP2 density itself) have the dipole moment here staying roughly constant and partial charges on the chlorine atoms undergoing a dramatic reduction. Overall, this means adding these two additional chlorine atoms is actually calculated to be unfavorable by nearly 1 kcal/mol, while experimentally it is *favorable* to add these. We are unclear why this is, as this discrepancy is robust across all the approaches we tried.

Thus, the key trends observed in ethanes (both with calculation and experiment) are that adding chlorines generally improves solvation, and improving solvent accessibility of chlorines also improves solvation. Increasing the dipole moment or polarity of the overall molecule tends to improve solvation, and there can be subtle effects depending on chlorine placement that play a role in this.

These trends are much harder to see, if they are present at all, in the data for the other two series. The experimental values may show some evidence for these for some of the compounds, but we see little sign of these trends in the calculated values. This is not necessarily a problem – it may be that as molecular complexity grows, other factors such as the details of the electron distribution (particularly, the pi systems present in these molecules) and structure become much more important than the number of chlorine atoms.

1.4.4 Observations for specific charge models and methods

As seen in Figure 1.1 and Figure 1 of the Supporting Information, most of the approaches we tried show rather similar trends. However, some specific observations are warranted. First, while our standard approach does well overall, it also exhibits a clear and incorrect trend for the dioxins. While the dioxin trend is incorrect across all our methods, the error is particularly pronounced in approaches using AM1-BCC partial charges (the standard, OEAnte-GAFF, OEAnte-Default, standard-newLJ, and standard-OPLS approaches).

Second, all the approaches we tested agree in computing hydration free energies for a number of compounds which are substantially too positive. This is moderated somewhat in the MP2/ExpSQ-Q case, where addition of virtual sites shifts the hydration free energies of many of these poly-chlorinated compounds to be somewhat more favorable, probably because these sites capture stronger C-Cl bond dipoles even while the overall polarity of the molecule remains weak. Still, even for this charge model free energies are still too positive on average. In some cases, there is some question about the quality of the experimental data [59]. This is denoted by large error bars (estimated by J. Peter Guthrie). The belief is that in these cases, if the experimental value is in error, it will typically be in error by being too negative. This could result because some of the specific experiments reported observing solubilities higher than the true value, which would yield hydration free energy estimates that are too negative [59]. Given the consistency across all of the approaches we tried, we are particularly concerned that this may be the case for 1,2,4,5-tetrachloro-3-

(3,4-dichlorophenyl)-benzene, 1,2,3,4-tetrachloro-5-(2,3,4,6-tetrachlorophenyl)-benzene, and the base dioxin, dibenzo-p-dioxin. Computed free energies are consistently too positive for some other members of the dioxin series as well, but the consistency and magnitude in those cases is not as pronounced. Given the consistency of this observation even across different Lennard-Jones parameters (including OPLS parameters) it seems likely the experimental data needs reevaluation.

Of the models we tested, the MP2/ExpSQ charges yielded the worst overall RMS error in absolute hydration free energies, but nearly the best overall correlation between predicted and experimental hydration free energies across all classes of compounds. As shown in Figure 1.1, the MP2/ExpSQ charges showed behavior similar to the AM1-BCC and MP2/SCRF charge sets: strong performance in predicting relative hydration free energies of ethane derivatives, prediction of the wrong trend in hydration free energies of the dioxin derivatives, and inability to predict any trends in the biphenyl derivatives. While the overall correlations with experimental hydration free energies were not high for any charge model, hydration free energies produced by all of the charge models are strongly correlated, particularly the MP2/SCRF and MP2/ExpSQ models as shown in Table 1.3.

As noted in Methods, after submission of our SAMPL results we also tried OPLS Lennard-Jones parameters. With AM1-BCC charges, these actually substantially improve the overall quality of the computed hydration free energies, increasing the R^2 and decreasing the average and RMS error relative to other charge models. As seen in Figure 1.1, most of the improvement comes from the biphenyls, where scatter is substantially reduced for a number of the compounds and error relative to experiment is reduced to less than 0.5 kcal/mol; the Kendall τ value is also just slightly increased for this set. This does not hold up, however, when OPLS Lennard-Jones parameters are used with MP2/SCRF partial charges – the average and RMS error again increase somewhat. (The R^2 and Kendall τ also decrease markedly, apparently chiefly because of a poor value for ethane, the single compound in this set with a hydration free energy that is very different from the rest.)

1.5 Discussion and conclusions

This test set, three different chemical series involving chlorinated derivatives of a common scaffold, represents an interesting and challenging set, especially given the relatively narrow range of experimental hydration free energies. While most of the approaches we tried gave reasonable agreement with experiment in terms of RMS and average errors (with RMS errors in the 1.4-2.0 kcal/mol range), this was actually somewhat less than satisfactory in this set given the narrow range of experimental values (with the RMS difference in experimental values also around 1.4 kcal/mol). Thus, essentially all approaches we tried do not do a good job at capturing experimental trends in hydration free energies, except for the ethane derivatives, where trends are typically captured reasonably well. Typically around 40% of pairs of compounds are correctly ranked (as measured with Kendall τ).

Our results suggest that, overall, current force fields still need substantial improvement in the area of accurately treating chlorinated derivatives of common scaffolds. While modifying the charge model (going from AM1-BCC to our MP2/SCRF approach) or switching Lennard-Jones parameters (from GAFF to OPLS) can yield modest improvement on this particular test set, even in the best cases, trends remain poorly captured, except for ethane derivatives. Perhaps these results are indicating that we can do no better for these compounds without going beyond an atom-centered partial charge model or atom types that are more dependent on the chemical environment, or perhaps Lennard-Jones parameters would need to be reoptimized in a charge-model-dependent way to improve agreement with experiment.

We also find, generally, that computed hydration free energies from our different approaches are strongly correlated (R^2 from 0.93 to 0.99 relative to MP2/SCRF, except for the MP2/OPLS set, with $R^2 = 0.75$, with much of the difference in this last case being due to substantial changes for ethane). In most of the sets we tried, the Lennard-Jones parameters are identical, which may be one major source of this correlation. However, even results with OPLS Lennard-Jones parameters remain fairly highly correlated. This is likely because the charge models themselves share numerous similarities that deserve examination. The strongest correlation in predicted hydration free energies occurs between charge sets fitted against MP2 calculations with similar basis sets (R^2=0.99); in one charge model, the aqueous solvent environment was mimicked by a dielectric boundary akin to Poisson-Boltzmann calculations, whereas in the other the aqueous solvent environment was represented by a time-averaged solvent charge density. As has been shown before, these two treatments actually create strikingly similar reaction fields within the solute [18], implying that the influence of each field on the solute charge density determined by MP2 calculations may have been very similar. Subsequent RESP fitting merely attempted to reproduce the field due to the solute charge density, not the field due to the solvent. No matter the charge model, the distribution of atomic partial charges was also similar: non-polarizable point charges placed at all atomic nuclei. Even in the AM1-BCC case, charges are parameterized to be like those from RESP fits. The process of RESP fitting naturally creates regions near the solute surface where the electrostatic potential is over- or under-estimated. While RESP fitting seeks to minimize mean square error in the potential, these regions of error might have been qualitatively similar across different models. Finally, another strong determinant of hydration free energies, the solvent model used during the thermodynamic calculations, was consistently set to TIP3P. No matter the charge model, water arrangements around the solute will be similar due to identical or at least extremely similar Lennard-Jones parameters. And the arrangement of water molecules will be a substantial contribution to the overall hydration free energy, especially since most of these compounds are not extremely polar.

Extension of the monopole basis set used in RESP fitting by massless virtual sites (in the MP2/ExpSQ-Q case) did substantially improve the fit to the quantum-mechanical electrostatic potential in regions accessible to solvent. In most cases, this extension of the charge model does not drastically change the hydration free energy. However, in poly-chlorinated biphenyl and dioxin compounds the virtual sites do appear to have a more pronounced effect in making the hydration free energies more favorable. None of these compounds showed strong overall dipoles, but the virtual sites permitted local dipoles along C-Cl bonds which models with only atomic nuclear charges could not fit. Although the solubilities of some of these compounds are uncertain, local dipoles may contribute to the unexpectedly high values reported in the literature.

Although the virtual site charge model afforded some success, it is noteworthy that all of the non-nuclear virtual sites took on positive values while driving the chlorine nuclear charges more negative. While numerically convenient for avoiding singularities in the simulations, this result underscores the artificial nature of the molecular mechanics charge model. Experience with modeling other compounds using the mdgx virtual site capabilities suggests that more virtual sites are of decreasing marginal value to improving the RESP fit. Complicated pi systems in the biphenyl and dioxin compounds likely cannot be modeled by simply adding virtual sites with fixed partial charges; although this treatment can create static dipole and quadrupole moments, polarization would likely be necessary to capture the physics of aromatic systems.

Despite the overall similarity of computed free energies across all of the different approaches, there are enough differences that when all approaches are wrong in a consistent way, it stands out. Particularly, there are three compounds identified above for which there is some question about the quality of the experimental data, and all of our tested approaches show substantial deviations from experiment. For these compounds, we believe our results suggest experimental follow-up is warranted.

Overall, agreement between calculated hydration free energies and experiment is decent in terms of average error, but there is considerable room for improvement in how classical molecular mechanics force fields capture hydration free energies for chlorinated derivatives in these three chemical series.

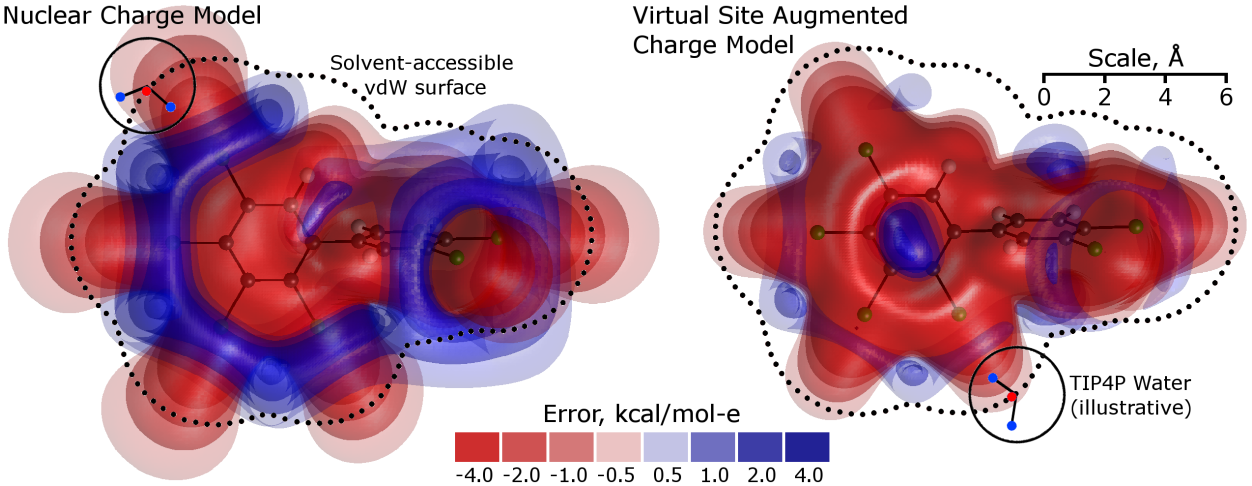

Nuclear Charge Model
Solvent-accessible vdW surface
Virtual Site Augmented Charge Model
Scale, Å
0
2
4
6
TIP4P Water (illustrative)
Error, kcal/mol-e
-4.0
-2.0
-1.0
-0.5
0.5
1.0
2.0
4.0

Figure 1.3: Errors in the electrostatic potential inherent in using point-charge models to approximate electron density derived from MP2 calculations. A set of point charges cannot perfectly reproduce the electrostatic field implied by quantum mechanical calculations. Some amount of error is due to the constraint that a set of fixed charges mimic the electrostatic potential for all molecular conformations, but a set of monopoles on atomic nuclei has more fundamental limitations. The two plots show differences between a quantum mechanical model and two fitted molecular mechanics approximations. The quantum mechanical model is a set of MP2 calculations performed in the context of the Gaussian SCRF continuum solvent model. The molecular mechanics approximation at left entails only charges at nuclear sites fitted with a RESP protocol. Significant errors in the electrostatic potential arise in solvent-accesible regions (the dotted line shows the boundary at which a TIP3P water molecule begins to experience a repulsive Lennard-Jones potential, and the illustrated water molecules show how TIP3P protons can sample regions even closer to the solute). The approximation at right adds additional sites not at nuclear positions (virtual sites) to better fit the potential near the chlorine atom as described in the text; the result in this case is a roughly two-fold reduction in the error over all space, most significantly near the solute surface. Including such virtual sites on all chlorinated compounds made a modest improvement to the computed hydration free energies.

Method	RMS (kcal/mol)	R^2 (kcal/mol)	Avg. err (kcal/mol)	AUE (kcal/mol)	Kendall τ
Standard	1.88 ± 0.07	0.64 ± 0.04	-1.34 ± 0.07	1.59 ± 0.06	0.43 ± 0.04
MP2/SCRF	1.59 ± 0.07	0.69 ± 0.04	-1.14 ± 0.06	1.22 ± 0.06	0.38 ± 0.04
MP2/ExpSQ	2.18 ± 0.07	0.70 ± 0.04	-1.92 ± 0.06	1.93 ± 0.06	0.40 ± 0.04
Standard-newLJ	1.90 ± 0.07	0.63 ± 0.04	-1.35 ± 0.07	1.59 ± 0.06	0.41 ± 0.04
OEAnte-GAFF	1.88 ± 0.07	0.64 ± 0.04	-1.34 ± 0.07	1.60 ± 0.05	0.42 ± 0.04
OEAnte-Default	2.02 ± 0.07	0.64 ± 0.04	-1.51 ± 0.07	1.71 ± 0.06	0.41 ± 0.04
Standard-OPLS	1.40 ± 0.05	0.74 ± 0.03	-0.87 ± 0.06	1.16 ± 0.04	0.46 ± 0.04
MP2/OPLS	1.48 ± 0.04	0.30 ± 0.07	-0.28 ± 0.08	1.26 ± 0.04	0.04 ± 0.04
MP2/ExpSQ-Q	1.33 ± 0.06	0.81 ± 0.03	-0.99 ± 0.05	1.06 ± 0.04	0.58 ± 0.03

Table 1.1: Statistics for different methods computing hydration free energies for SAMPL, with methods abbreviated as described in the text. We compare the RMS error, the Pearson correlation coefficient (R^2), the average error, the average unsigned (absolute) error (AUE), and the Kendall τ value.

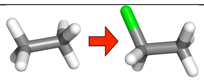
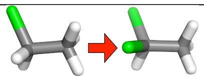
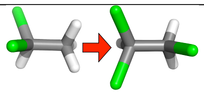
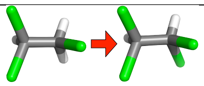
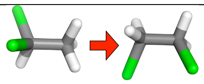
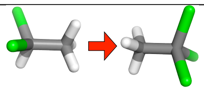
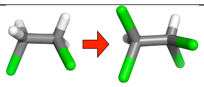
	Transformation	$\Delta\Delta G_{expt}$	$\Delta\Delta G_{calc}$	$\Delta\Delta G_{calc,e}$	$\Delta\Delta G_{calc,LJ}$
(a)		-2.25 ± 0.14	-2.74 ± 0.01	-2.31 ± 0.01	-0.42 ± 0.01
(b)		-0.49 ± 0.14	-0.32 ± 0.01	-0.08 ± 0.01	-0.24 ± 0.01
(c)		-1.09 ± 0.14	-0.34 ± 0.05	-0.06 ± 0.01	-0.29 ± 0.04
(d)		-0.40 ± 0.14	0.21 ± 0.05	0.40 ± 0.01	-0.20 ± 0.03
(e)		-0.92 ± 0.14	-1.00 ± 0.01	-0.81 ± 0.01	-0.19 ± 0.01
(f)		0.62 ± 0.14	0.88 ± 0.01	1.0 ± 0.01	-0.14 ± 0.01
(g)		-0.57 ± 0.14	0.87 ± 0.02	1.16 ± 0.01	-0.29 ± 0.02

Table 1.2: Changes in hydration free energy on adding a chlorine, or moving chlorines, for selected pairs of ethane derivatives. Shown are calculated and experimental hydration free energy changes, as well as the portion of the calculated change due to electrostatics ($\Delta\Delta G_{calc,e}$) and the portion due to changing Lennard-Jones interactions ($\Delta\Delta G_{calc,LJ}$). All calculated values are from the MP2/SCRF set.

Method	R^2 with MP2/SCRF
Standard	0.96±0.01
MP2/ExpSQ	0.99±0.01
Standard-newLJ	0.96±0.01
OEAnte-GAFF	0.95±0.01
OEAnte-Default	0.93±0.01
Standard-OPLS	0.94±0.01
MP2/OPLS	0.75±0.03
MP2/ExpSQ-Q	0.95±0.01

Table 1.3: Correlation coefficients (R^2) between MP2/SCRF results and results obtained with various other approaches.

Chapter 2

Blind prediction of HIV integrase binding from the SAMPL4 challenge

2.1 Abstract

Here, we give an overview of the protein-ligand binding portion of the SAMPL4 challenge, which focused on predicting binding of HIV integrase inhibitors in the catalytic core domain. The challenge encompassed three components – a small "virtual screening" challenge, a binding mode prediction component, and a small affinity prediction component. Here, we give summary results and statistics concerning the performance of all submissions at each of these challenges. Virtual screening was particularly challenging here in part because, in contrast to more typical virtual screening test sets, the inactive compounds were tested because they were thought to be likely binders, so only the very top predictions performed significantly better than random. Pose prediction was also quite challenging, in part because inhibitors in the set bind to three different sites, so even identifying the correct binding site was challenging. Still, the best methods managed low RMSD predictions in many cases.

Here, we give an overview of results, highlight some features of methods which worked particularly well, and refer the interested reader to papers in this issue which describe specific submissions for additional details.

2.2 Introduction

Accurate protein-ligand binding predictions could impact many areas of science. An ideal computational method which could quickly and reliably predict binding free energies and bound structures for small molecules of interest to arbitrary receptors would have far reaching applications, including in virtual screening, drug lead optimization, and even further afield, to help enzyme design, systems biology, and in a variety of other applications. However, most systematic tests of methods for predicting binding strengths and binding modes indicate that these still need substantial improvement to be of routine use in discovery applications. While methods can be improved based on existing experimental data, methodological improvements need to be tested in a predictive setting to determine how well they work prospectively, and especially so for methods involving empirical parameters which are tuned to fit previously known values. Thus, we need recurring prediction challenges to help test and advance computational methods. The Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenge we discuss here provides one such test.

2.2.1 SAMPL protein-ligand binding background

The SAMPL challenge focuses on testing computational methods for predicting thermodynamic properties of small drug-like or fragment-like molecules, including solvation free energies, host-guest binding affinities, and protein-ligand binding. The challenge started informally in 2007 at one of OpenEye Software's Customers, Users, and Programmers (CUP) meetings[130], and then was formalized as the SAMPL challenge beginning in 2008. Here, we discuss the results of the protein-ligand binding component of SAMPL4, the 4th iteration of the SAMPL challenge, which took place in 2013.

Protein-ligand binding has not been a feature of every SAMPL challenge, featuring previously only in SAMPL1 and SAMPL3. SAMPL1 included a pose prediction test on kinases, which proved extremely challenging. The system which proved most interesting for analysis was JNK3 kinase, where the best performing predictions were from two participants who used software-assisted visual modeling to generate and select poses. Essentially, pose predictions were generated and filtered using expert knowledge of related ligands or related systems. The two experts applying this strategy substantially outperformed all pure algorithmic approaches [159]. Affinity prediction results in some cases were reasonable, however[165]. The SAMPL3 protein-ligand challenge involved predicting binding of a series of fragments to trypsin [129], and a number of groups participated[97, 166, 98, 165, 4], in some cases achieving rather good enrichment for screening [166, 165] and good correlations between predicted binding strength and measured affinity[166], though the test was still challenging[98].

The current SAMPL4 challenge focused on predicting binding of a series of ligands to multiple binding sites in HIV integrase.

2.2.2 HIV integrase background

According to the World Health Organizations data (`http://UNAIDS.org`), over 33.3 million people are currently living with an HIV infection. Approximately 2.3 to 2.8 million people become infected with HIV annually, and 1.7 million people die from HIV-related causes each year. Throughout the AIDS epidemic, over 32 million people have died of HIV-related causes, which makes HIV the deadliest virus plaguing humanity.

HIV integrase and the drugs that target it

HIV integrase (IN) is one of three virally encoded enzymes. It performs two distinct catalytic functions called “3’ processing” (which cleaves two nucleotides off of the end of the viral cDNA in a sequence-specific manner to generate reactive $\mathrm{CA_{OH}}$-3’ termini) and the “strand transfer reaction” (which covalently attaches, or integrates, the cleaved viral cDNA into human genomic DNA, in a non-sequence-specific manner). Two drugs that target the active site of IN have been approved by the FDA for the treatment of HIV/AIDS: Raltegravir was approved in 2007, and Elvitegravir was approved in 2012[142]. These two drugs are called INSTIs (for Integrase Strand Transfer Inhibitors). A third INSTI, Dolutegravir, is currently in late-phase clinical trials[142].

HIV integrase is an enzyme that is part of a large family of recombinases that all contain the “DDE” motif (or D,D-35-E motif) within the active site. The two Asps and one Glu are used to chelate two magnesium ions, using monodentate interactions between each carboxylate group and a magnesium[140]. This active site region is where the 3’ processing and strand transfer reactions occur. One monomer of HIV IN contains three different domains: the N-terminal domain (NTD), catalytic core domain (CCD), and the C-terminal domain (CTD) (Figure 2.1). When performing catalysis (or when bound to DNA immediately before catalysis occurs), IN is a tetramer (i.e., a dimer of three-domain dimers). The NTD is an HH-CC zinc-binding domain (for the His,His-Cys,Cys motif that chelates the Zn). The CTD displays the SH3 fold and binds DNA non-specifically (to likely help position, or scaffold, the DNA and direct it towards a CCD). The CCD displays the RnaseH fold, and two monomers of the CCD form a spherical dimer. The CCD dimer contains two active site regions (i.e., one active site per monomer), which is where the advanced INSTIs all bind (i.e., they bind to the complex of the CCD with DNA). But in the full 3-domain tetramer of HIV IN bound to both viral cDNA and human genomic DNA, it is likely that only one active site per CCD dimer is involved in catalysis (due to geometric constraints). All of the crystal structures of

HIV IN used in this challenge contained dimers of only the CCD. Although there are many crystal structures of HIV IN, the only structures available in the PDB contain only one or 2 domains of the full 3 domain monomer of IN, and none of the HIV IN crystal structures include any DNA. However, a similar recombinase from Prototype Foamy Virus, called PFV integrase, was recently crystallized in many different complexes by Peter Cherepanov, et al. These PFV IN crystal structures often contain DNA and one of the three aforementioned advanced INSTIs (see Figure 2.1)[106, 63, 62, 64, 114].

Although the three advanced INSTIs were only recently developed, multi-drug-resistant mutants against which these inhibitors lose their potency have already appeared in clinical settings[141, 142, 176, 53, 1]. There are three main, independent pathways resulting in INSTI resistance, which involve mutations at positions Tyr143, Asn155, and Gln148[176], all of which are within the catalytic active site region. Mutations at these positions, especially when combined with additional secondary mutations at other positions, cause extensive cross-resistance to both Raltegravir and Elvitegravir, and mutations involving Gln148 significantly decrease the susceptibility of HIV to all three advanced INSTIs[176, 53, 1]. The fact that HIV can quickly evolve drug resistance against these strand transfer active site inhibitors of IN highlights the urgent need to discover and develop new classes of drugs that bind to different sites and display new mechanisms of action.

Utility of allosteric inhibitors

Combinations of two different classes of inhibitors that act on the same enzyme have been shown to inhibit broad panels of many different multi-drug-resistant mutants and to also decrease the probability of the emergence of new drug-resistant mutants, as exemplified by the combination of an active site inhibitor and an allosteric inhibitor of Bcr-Abl, a kinase target for cancer chemotherapy[80, 190]. This also appears to be the case for HIV treatment. When a Nucleoside Reverse Transcriptase Inhibitor (NRTI; NRTIs target the active site of

HIV reverse transcriptase, or RT) is combined with an allosteric Non-Nucleoside Reverse Transcriptase Inhibitor (NNRTI; NNRTIs target a non-active site region of HIV RT), the evolution of drug resistance to both classes of drugs is impeded[29].

LEDGF inhibitors

One new class of inhibitors that seem particularly promising in the fight against AIDS are the subset of ALLINIs (for ALLosteric INtegrase Inhibitors) called LEDGINs[26] , which bind to the LEDGF site at the dimer interface of the catalytic core domain[95, 39]. LEDGF (for Lens Epithilial-Derived Growth Factor) is a human protein that HIV exploits: when IN interacts with LEDGF/p75, it guides the integration of the viral genome into the regions of our chromosomes where the actively expressed genes are located[20], which increases the probability of the subsequent production of viral proteins that can then help spread the infection. The LEDGINs use a carboxylate group to mimic a key interaction that LEDGF utilizes to bind to the backbone amino groups of Glu170 and His171, which are located in the LEDGF site of IN[95]. When LEDGINs bind the LEDGF site, they promote and stabilize higher-order multimers of IN and inhibit the catalytic process[92, 169, 25, 85].

Before the challenge began, the participants were informed that most of the SAMPL4 compounds were known to bind to (at least) the LEDGF site of IN, but some of the compounds were known to bind to at least one of the two additional allosteric sites of IN, which were referred to as the "FBP" site (for Fragment Binding Pocket) and the "Y3" site (see Figure 2.1). Like the LEDGF site, the FBP site is also located at the dimer interface of the catalytic core domain (CCD) of IN. There are two LEDGF sites per IN CCD dimer, two FBP sites per IN CCD dimer, and also two Y3 sites per IN CCD dimer. But the Y3 site is entirely contained within each monomer of the core domain and is located underneath the very flexible 140s loop (i.e., Gly140-Gly149). The top of the 140s loop flanks the active site region, and the composition, conformation, and flexibility of the 140s loop is known to be critical to IN activity[57, 140, 35] . Participants in the pose prediction challenge were given the hint that, if they were concerned about trying to predict the binding site, they might wish to focus their efforts on the LEDGF site, though most chose not to do so. This could have led to successful binding mode predictions in 52 of 55 cases considered.

2.3 SAMPL challenge preparation and logistics

The experimental data for the IN portion of SAMPL4 is described in detail elsewhere in this issue[137]. It includes a set of inactive compounds which were not observed to bind via both crystallography and surface plasmon resonance (SPR), and a set of actives; together, these were used for the virtual screening component of SAMPL4. Additionally, crystal structures for some 57 of the actives were used for the pose prediction challenge. Accurate affinities were measured via SPR for 8 of these compounds, and these were used for the affinity prediction challenge.

For each portion of the challenge, participants were provided with a PDF of introductory material on the system prepared by Thomas S. Peat, which included a brief overview of the biological relevance, the different binding sites, and some references to previous published work from the same discovery project. This PDF is provided in the Supporting Information. In addition to this PDF, participants in each individual component received a further set of calculation inputs which will be described below.

The integrase portion of SAMPL4 was staged, so that participants must either complete or opt-out of virtual screening before going on to pose prediction, and complete or opt-out of pose prediction before going on to affinity prediction. This was done because inputs for the subsequent portions of the challenge would reveal all or part of the results from the earlier challenge components. In some cases, participants opted to conduct the whole challenge using *only* the inputs for the virtual screening challenge, and thus had no information about the identities of actual binders and/or structures when working on the affinity prediction and pose prediction challenges. This was primarily the case for submission IDs 535-540.

The SAMPL4 challenge was advertised via the SAMPL website (http://sampl.eyesopen.com) and e-mails to past participants, others in the field, and the computational chemistry list (CCL), beginning in January, 2013. The virtual screening portion of the challenge was made available via the SAMPL website April 1, 2013, and participants moved on to the other components once their screening results were submitted, or once they opted out. Submissions for all challenge components were due Friday, Aug. 16. The challenge wrapped up with the SAMPL4 workshop on Sept. 20 at Stanford University. Submissions were allowed to be anonymous, though we[1] received only three anonymous submissions from this portion of the challenge. Because of this, however, we typically refer to submissions by their submission ID (a three digit number) rather than by the authors' names.

2.3.1 Pre-challenge preparation

The challenge organizers were provided with three main inputs to prepare the SAMPL4 challenge. First, we received a disk with raw crystallography data and refined structures for the majority of the compounds which were crystallized. Second, we received a spreadsheet describing the active compounds, with SMILES strings, 2D structures, information about the density, and the location of the data on the disk. Third, we received a document containing images of the chemical structures of many inactive compounds. Fourth, we received a list of the molecules for which affinities were being measured precisely via SPR. Our pre-challenge preparation mainly involved turning this information into suitable inputs for predictions, and checking the data. Here, we used OpenEye unified Python toolkits[131] unless otherwise

[1] The SAMPL4 challenge was designed, run and evaluated by the Mobley lab with some help from Kim Branson, so when this report uses the word "we" to refer to an action relating to challenge design, logistics, and analysis, it refers to these authors – specifically, Mobley, Branson, Su, Lim, Wymer, and Liu

noted.

Preparing inactives

For the list of non-binders, since we had only compound identifiers and images of the 2D structures, we re-drew 2D structures of all of the non-binding compounds in Marvin Sketch [111] and then stored SMILES of these which were subsequently canonicalized and turned into 3D structures using the OpenEye toolkits[131] and Omega[66, 65]. Since this step involved manually drawing the structures, all structures drawn were inspected by two different people to check for accuracy.

Preparing actives

We also needed SMILES strings and 3D structures for all of the binders. SMILES strings were available both in the spreadsheet we were provided and on the disk, but these were not always consistent, and typically omitted stereochemistry information. We found that the most reliable route to getting this information was to pull the 3D ligand structures from the protein structures we were provided, then add protons and perceive stereochemistry information based on these structures. However, strain or other issues in the structures on occasion resulted in incorrect assignment of stereochemistry.

To deal with incorrect assignment of stereochemistry, we used OpenEye's Flipper module to enumerate all stereoisomers for each ligand, and with the Shape toolkit overlaid these onto the ligand structures pulled from the refined PDB files, automatically selecting the best-scoring shape overlay as the correct stereoisomer for cases with high shape similarity. Any alternate stereoisomer case where the shape Tanimoto score was within 0.1 of the best scoring shape overlay was flagged for additional manual inspection, although ultimately all structures were inspected manually. Based on manual examination of the shape overlays and electron densities in cases where there was any ambiguity, we concluded that the automatically assigned stereochemistry information was correct in every case except AVX17587, 38673, 38741, 38742, 38747, 38748, 38749, 38782, 38789, 101124, and GL5243-84. This seemed primarily to be because of poor-quality shape overlays in these cases, possibly due to ligand strain. Once we finished applying this procedure, we saved 3D structures of the correct stereoisomer of every ligand, as well as the isomeric SMILES string specifying stereochemistry information. In some cases our shape overlay work here actually resulted in a re-evaluation and potentially a re-refinement of the crystal structure, as discussed elsewhere[137].

Stereoisomer enumeration

In general, chiral compounds were tested as a mix of stereoisomers, so treating isomers as distinct compounds provides an opportunity to expand the list of inactive compounds. This is especially true for the inactive compounds, but even for the active compounds, if a given stereoisomer is not observed to bind, it means either that it does not bind, or it is much weaker than the stereoisomer which is observed to bind. Thus, for all compounds we enumerated all stereoisomers using Flipper and assigned them an isomer ID which was added to their ID. For example, for AVX38670, with two stereoisomers, these were labeled AVX38670_0 and AVX38670_1 and treated separately for the virtual screening and (when applicable) pose prediction challenges. The issue of whether or not to treat alternate (apparently non-binding) stereoisomers of actives as inactives will be discussed further below.

After generating or reading in isomeric SMILES strings for all compounds, we also cross-checked for duplicate compounds under the same or different identifiers and removed a number of such duplicates. In the OpenEye toolkits, there is a 1:1 correspondence between an isomeric SMILES string and a particular compound in its standard representation, so we expected that this would catch all duplicates. However, because of differences in how bonding was assigned prior to generating isomeric SMILES strings, some SMILES strings were generated from the Kekulé representation of molecules and some were not. These forms result in different isomeric SMILES strings, so some duplicates remained when we conducted the challenge and were only removed when we discovered this in post-analysis, as discussed below.

Protonation state assignment

By default, protonation/tautomer states for provided 3D structures for all compounds were assigned via the OpenEye toolkits using their "neutral pH model" predictor, though we did some additional investigation for pose prediction and affinity prediction, as noted below.

Molecular dynamics re-refinement

In the process of preparing for SAMPL, several structures were re-refined and in several cases resulted in substantial changes. We were concerned that we might miss other problem cases, and sought an automated procedure to identify cases where the binding mode might be questionable. Therefore, we took all refined structures and simulated them in the AMBER99SB-ILDN protein force field [101] with the AMBER GAFF[179] in GROMACS 4.6.2 for 110 ps of equilibration and another 100 ps of production, using protein protonation states assigned by MCCE. Equilibration was done gradually releasing restraints on the protein+ligand. Following this, we monitored RMSD over the course of the short production simulations and looked for cases where the ligand moved substantially away from its starting binding mode, by more than 3 Å RMSD. This flagged several cases as potentially problematic – AVX17558, AVX38749, and AVX38747. All three have a somewhat-floppy alkyl tail which in at least two of the cases has fairly poor density, which may be part of the issue. Re-refinement from our final structures from MD did not result in substantial improvement. Still, to us this suggests that closer scrutiny of these three may still be warranted. Particularly, in AVX17558 and AVX38747, there is some question as to the chirality. For AVX17558, there is some evidence in the density that both stereoisomers bind[138], while for AVX38747, it is not completely clear which stereoisomer fits the density best[138]. The remaining cases remained quite close to the crystallographic structure.

2.3.2 Virtual screening

In addition to the IN background PDF noted above, virtual screening participants were provided with a README file, a template for submitting their predictions, isomeric SMILES strings and 3D structures (in MOL2 and SDF format) for all stereoisomers of all compounds, and a reference protein/ligand structure in the form of a `3NF8_reference.pdb` file – essentially, the PDB 3NF8 structure, aligned to the frame of reference we had chosen for the challenge. This 3NF8 structure was selected in part because it contains a bound ligand from the series studied here, and in part because this ligand is observed in all six binding sites (both copies of the LEDGF, Y3, and FBP sites). The README file contained information on what they were to submit, notes about the reference structure and the locations of the three sites, and a substantial hint – that "many (though by no means all) of the ligands bind in the LEDGF site, so if you like, you can focus on just that site and still do relatively well." We also included a disclaimer that the ligand protonation/tautomer states are provided "as is" and participants might wish to investigate these on their own. Submissions included a rank for each compound, a field indicating whether or not it was predicted to bind ("yes" or "no") and a confidence level ranging from 1 (low confidence) to 5 (very confident). These files are provided in the Supporting Information.

After the challenge, we found some issues with duplicate or incorrect compounds included in the virtual screening set. Specifically, we had to remove AVX17684m (or AVX17684-mod) because it was present in only some of the files which were distributed, and AVX17268_1 because it was incorrect. And only one member of each given set of duplicates was considered in analysis. Duplicates/replicates included (AVX17556, AVX17561, and GL5243-84), (AVX17557 and AVX17587), (AVX101125 and AVX62777), (AVX17285 and AVX16980), and (AVX17557 and AVX17587).

2.3.3 Pose prediction

In addition to the IN background PDF, pose prediction participants were provided with a README file, the reference structure described above, and SMILES strings and 3D structures (in MOL2 and SDF format) for all ligands, as in the case of the virtual screening challenge. The main differences here were that in this case, the compound list included only active compounds, and additional information in the README file. Particularly, the README file additionally added some additional pointers concerning protonation/tautomerization states. For this challenge portion, we used Epik, from Schrö-dinger, to enumerate possible protonation and tautomer states, and cross-compared the Epik predictions with those from OpenEye's QuacPac[131]. As a result, we highlighted compounds AVX17715, AVX58741, AVX38779-38789, and AVX-101118 to 101119 as having possible uncertainty in their protonation states, and GL5243-102 as having two possible tautomers on its five-membered ring, so these notes were provided in the README. The input files are provided in the Supporting Information.

Depending on the nature of their method, participants submitted either a 3D structure of the ligand in its predicted binding mode (relative to the 3NF8_reference.pdb structure provided), omitting the protein; or a 3D structure of the ligand-protein complex. In this challenge, our analysis focused only on the predicted binding modes relative to a static structure, so in cases where the full protein structure was submitted (i.e. for flexible protein methods), we scored binding modes based on an alignment onto the static reference structure.

For pose prediction, participants received SMILES strings for 58 compounds but 3D structures for 65 compounds because of a scripting error which resulted in some extra isomers being included in the 3D structures directory. So participants should have predicted binding modes for 58 compounds. However, several additional compounds had to be removed prior to analysis. Specifically, AVX17680 was removed because participants were provided with the wrong SMILES string and 3D structure because of a scripting error. Additionally, AVX101121 had a discrepancy between its SMILES string and its structure (differing by a methyl) apparently due to confusion about the original identity of the compound in the experiments, so this was removed prior to analysis. A similar thing happened with AVX-17543 on the computational end – participants were given an incorrect ligand SMILES and structure, and this had to be removed prior to analysis. Finally, AVX-17557 and AVX-17587 are actually the same compound, prepared as different salts. Thus the final number of compounds analyzed was 54.

2.3.4 Affinity prediction

In addition to the IN background PDF, affinity prediction participants were provided with 3D structures of all 8 ligands (in MOL2 and SDF format), a README file, the refined crystal structures PDB format, MTZ format density files for the crystal structures, shape overlays of the ligands onto the crystallographic ligands (generated by the OpenEye Shape toolkit[131]), the refined crystal structure and ligand for the compound from the 3ZSQ structure, which was used as the control compound in SPR, and a text file template for submissions, which contained fields for the compound ID, the predicted binding free energy, the predicted statistical uncertainty, and the predicted model uncertainty. In this case, the README file highlighted minor issues with the electron density for AVX-17557 and for the aliphatic amino in AVX-38780, and an alternate rotamer for Leu102 in AVX40811 and AVX40812, as well as uncertainty in the protonation state of AVX38780.

2.4 SAMPL Analysis Methods

In general, analysis was done using OpenEye's Python toolkits for working with molecules and structures, and Python/NumPy for numerical data. Matplotlib was used for plots.

2.4.1 Virtual screening

Virtual screening performance was analyzed by a variety of relatively standard metrics, including area under the curve (AUC) and enrichment factor at 10% of the database screened (EF10), as well as the newer Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC)[168]. We also made enrichment and ROC plots for all submissions. These were done using our own Python implementation of the underlying routines.

2.4.2 Pose prediction

Here, we focused primarily on judging pose predictions via RMSD. We used two different evaluation schemes depending on how we handled cases with multiple copies of the ligand bound. Since IN is a dimer, there are two essentially symmetric copies of each binding site, for a total of 6 binding sites. These "symmetric" sites exhibit non-crystallographic symmetry (sequence symmetry) and are in some cases not quite symmetric. Typically, they in fact were refined separately. This introduced some complexities for judging pose prediction. Even if a ligand only occupied the LEDGF site, and a participant only predicted one binding mode, two RMSD scores were possible depending on how the prediction was superimposed onto the crystallographic structure. To compute both values, we rotate the crystal structure to the alternate possible alignment onto the reference structure (thus handling the non-crystallographic symmetry). Then, we compute the RMSD based on both the original alignment and the new alignment, and retain the best value as the score for this submission.

This scenario of only a single predicted binding mode applied to the majority of submissions, though a minority of participants predicted multiple binding modes for some ligands, and a minority of ligands bound in other binding sites or exhibited multiple site binding. In these cases, additional RMSD values were possible. For example, if a participant predicted a ligand to bind to the LEDGF site, and actual binding was observed in both LEDGF sites and the Y3 site, we would obtain four different RMSD values. To handle this ambiguity, we chose to use two different scoring schemes, which we call "by ligand" and "by pose". In the "by ligand" scheme, we choose each submission's best RMSD value for each ligand, resulting in a total of 54 RMSD values. In the "by pose" scheme, each experimental binding site and mode (LEDGF, Y3, FBP) is scored separately and the best RMSD value is retained for each, resulting in a total of 112 RMSD values. Since most participants predicted only one binding mode per compound, this latter scheme penalizes submissions which miss binding to the additional sites in the case of multiple-site binding, while the former does not.

Our analysis here focuses only on scoring the best RMSD value of each submission for each ligand or pose. In general one might also be interested in knowing the worst RMSD. But since all but two participants here submitted only a single binding mode for each compound, the best and worst values are essentially identical for most submissions here.

To evaluate RMSD scores, we used a maximal common substructure search to match predicted ligand binding modes onto the crystallographic binding modes since this matching was not always obvious. Particularly, submissions used a variety of file formats, and not all submissions included ligand hydrogen atoms. Some submissions also altered atom naming conventions, meaning that the most straightforward approach of simply matching atoms by their names would not always work. Additionally, some ligands had internal symmetries (for example, a symmetric, rotatable ring) and participants should not be penalized for flipping symmetric groups. So, for each ligand or pose considered, we evaluated multiple maximum common substructure matches using the OpenEye Python toolkits, and took the match yielding the lowest RMSD. This approach simultaneously handled the issue of internal symmetry, together with variations in atom naming and protonation state.

In some cases, portions of ligands were relatively flexible and had only weak electron density. For example, a number of ligands had a floppy alkyl tail which was relatively poorly resolved but still included in the refined structures. We wanted to avoid penalizing participants for predictions which did not fit the model well in regions of weak density. Therefore, we manually inspected the electron density for all ligands and built a list of ligand heavy atoms which did not fall within the $2F_o - F_c$ density when contoured at 1σ. This included C54 and N57 for AVX-17557; C42, C48, and N51 for AVX-17558; C18 for AVX17684m; C18 and O29 for AVX38672; N30, N31, C24 and C6 for AVX38741; N23, O30, C17, O27 and O25 for AVX38742; C25, C26, and O28 for AVX38743; C22 for AVX38747; C14, O31, C22, C25, and C26 for AVX38748; O32 for AVX38749; O1, O25, and O26 for AVX101140; and C20, C21, C22, and C23 for GL5243-84. These atoms were excluded from RMSD calculation, so in these cases only the portions of the ligands which did have good electron density were counted for scoring. In this case, since most submissions did relatively poorly at predicting binding modes, this consideration did not substantially alter RMSD values. However, we believe this procedure is in general good practice to avoid a scenario where one method appears better than another simply because it gives binding modes more consistent with those from refinement, even when there is no difference in how well they fit the electron density.

We had originally planned to also calculate the diffraction-component precision index (DPI), and thus the coordinate error, for each of the structures[6]. This would provide a mechanism to compute the best achievable RMSD values. For example, two methods can be considered equally good whenever they yield the lowest RMSD which can be obtained given the coordinate error. Or, to put it another way, RMSD comparisons are useful only for RMSD values above the fundamental limit imposed by the precision in the coordinates. Experimental structures with very precise coordinates can permit RMSD comparisons down to very low values, while less precise structures provide less information about which predictions are the most accurate. This could be dealt with in analysis by assigning a DPI-adjusted RMSD to any submission which coincidentally obtained an RMSD value lower than the best possible value expected given the coordinate error. In general, we believe this approach is the correct one to take in comparing binding mode predictions by RMSD. However, we ran out of time to conduct this analysis prior to SAMPL, and here, so many binding modes proved very difficult to predict that small adjustments to the RMSD values for a few submissions on a few ligands would not have substantially affected the overall analysis.

One other metric commonly used to assess binding mode prediction is the fraction of ligands correctly predicted. However, "correct" is typically defined with respect to an arbitrary cutoff – for example, ligands predicted better than 2 Å RMSD might be said to be correctly predicted, while another practitioner might use a different cutoff. To avoid this ambiguity here, for each submission, we plotted the fraction of ligands correctly predicted as a function of RMSD cutoff x and evaluated the area under the curve (AUC). These plots and the AUC were provided to participants and are discussed below. In general, a method will predict no binding modes correctly at a cutoff of 0 Å RMSD, and all binding modes correctly at a cutoff larger than the size of the receptor, with the fraction correct varying in between these. A reasonable AUC can be achieved in multiple ways – for example, by having many very accurate predictions but also many very wrong predictions, or by having all predictions achieve modest accuracy.

As noted above in Section 2.3.1, there may still be questions about the true ligand binding mode or bound structure in a handful of cases. However, the set is large enough that these cases do not substantially affect the conclusions of the analysis here, and so our overall analysis includes these cases.

It is worth noting that the vast majority of the ligands in this series bind exclusively in the LEDGF site, with a smaller number exhibiting multiple site binding, and a few binding only in alternate sites. Specifically, AVX-15988, AVX-17389, and AVX-17679 bind FBP exclusively; and AVX-17631 binds both the LEDGF site and the FBP. pC2-A03 binds just one of the Y3 sites, and AVX-17258 and AVX-101140 bind both Y3 and the LEDGF site. A few structures are annotated with other possibilities – AVX-17260 is noted to have some density in the FBP, but is only modeled in the LEDGF site; and AVX-17285 is suggested to perhaps have multiple conformations but only one is modeled.

This analysis focused on ligand binding mode prediction essentially in the absence of protein motion, and even in cases where participants used a flexible protein and submitted a protein structure along with the ligand binding mode, this was used only to align their protein structure onto the reference structure used for judging. We made this choice primarily because protein motion here was quite minor, with side-chain rearrangements only appearing in a handful of cases (for example, LEU102 for AVX-17377, AVX-40811, and AVX-40812) and GLU157 in pC2-A03 and AVX-17377), and more substantial protein motion appeared to in general be absent. Thus it seemed appropriate to focus this SAMPL primarily on ligand binding mode prediction within an essentially static protein structure.

Because binding site prediction was a substantial challenge which was here convoluted with binding mode prediction, we recomputed all metrics for each submission for the fraction of poses which were placed “in” the correct binding site. For the purposes of this analysis, we considered compounds in the correct binding site when they were placed so that the predicted center-of-mass (COM) location is nearer the COM of a ligand in that binding site in the 3NF8 structure than the COM of the ligand in any other binding site. In other words, each ligand was considered to be in the binding site its center-of-mass was nearest to.

We also computed an interaction fingerprint metric to look at whether predicted binding modes made the correct contacts with the protein. Interaction fingerprints were computed using Van der Waals (VdW) interactions from the DOCK 3.5 scoring function[42], and the hydrogen bonding term from SCORE[182]. For each atom in the protein the VdW and hydrogen bonding interactions were checked against every atom in the ligand. A bit string was constructed for each atom in the protein. Protein atoms with a favorable VdW or hydrogen bonding interactions had their bits set to 1 otherwise 0. A bit string was calculated for the crystallographic ligand coordinates (the reference string) and the docked poses. The Tanimoto coefficient was used to assess the similarity in protein contacts between the reference and docked poses.

Several minor changes were made to structures after the SAMPL challenge and all analysis was completed, during the process of deposition to the PDB. Because these changes were made at such a late date, when many SAMPL manuscripts were already in review and/or accepted for publication, our analysis was left as is and these updates are only noted here. Specifically, further work on AVX-38743 determined that a mixed regio-isomer is actually a better fit to the density, as seen in the final structure (PDB 4CF9). And for AVX-38741, it was determined that a ring which had been thought to have formed within the molecule in fact did not form, altering the compound identity (PDB 4CF8).

2.4.3 Affinity prediction

Of the actual binders observed here by crystallography, only a small number were strong enough to obtain accurate affinities via surface plasmon resonance (SPR). Affinities measured by SPR were provided for 8 compounds by Tom Peat and collaborators[137]. These were provided as K_d values with uncertainties and converted to ΔG° for analysis. A couple of additional compounds were also available, but due to questions about the stoichiometry of binding and other issues these were excluded from the analysis. Final affinities are all fairly weak, spanning from 200 to $1460 \mu M$, unfortunately giving a rather narrow range of binding free energies.

All submissions were analyzed by a variety of standard metrics, including average error, average unsigned error, RMS error, Pearson correlation coefficient (R), and Kendall tau, as well as the slope of a best linear fit of calculated to predicted values. Additionally, we compared the median Kullback-Leibler (KL) divergence for all methods, adjusted to avoid penalizing for predicted uncertainties that are smaller than the experimental error when the calculated value is close to the experimental value, as discussed in more detail elsewhere in this issue[127]. Because KL divergences are difficult to average when performance is poor, we also looked at the expected loss, given by $L = \left\langle 1 - e^{-(KL)} \right\rangle$ where KL is the KL divergence[127].

We also examined one additional metric, what we call the error slope, which evaluates how well submissions predicted uncertainties. This looks at the fraction of experimental values (resampled with noise drawn from the experimental distribution) falling within a given multiple of a submissions assigned statistical uncertainty, and compares it to the fraction expected (a Q-Q plot), as discussed elsewhere in this issue[127]. A line is fit to this, and a slope of 1 corresponds to accurate uncertainty estimation; a slope higher than 1 means uncertainty estimates were too high on average, and a slope lower than 1 means uncertainty estimates were too low on average.

2.4.4 Error analysis

For all sections of the challenge, we computed uncertainty estimates in all numerical values as the standard deviation measured over a bootstrapping procedure as explained in more detail elsewhere in this issue[127]. Some additional detail is warranted for the virtual screening analysis, where bootstrapping consisted of constructing “new” datasets by selecting a new set of compounds of the same length at random from the original set, with replacement, and pairing these with the corresponding predicted values. This new set typically contained multiple entries of some compounds and omitted others, allowing assessment of the dependence of the computed results on the set. As usual, the uncertainty was reported as the standard deviation over 1000 bootstrap trials.

2.5 Integrase Screening Results

2.5.1 SAMPL analysis focused on the full set

For the binding prediction portion of the challenge, we received 26 submissions from nine different research groups. Overall statistics for these are shown in Table 2.1 and Figure 2.2. We also show statistics for two control or null models, 007 and 008, described below. In general, this portion of the challenge was extremely challenging, and even the best methods enriched actives only slightly better than random over the entire set of 305 compounds (with 249 non-binders)[2]. We attribute this to several factors. First, participants did not know the actual binding site, increasing the potential for false positives. Second, the inactive compounds here are available precisely because they were thought to be good candidate binders and therefore were tested experimentally. That is, they are part of the same series as the active compounds and resemble the active compounds in essentially every respect. Third, many (116) of the inactives are in fact alternate stereoisomers of active compounds, further increasing their resemblance to actives. The challenging nature of this test can be observed by noting that only 5 submissions (submission IDs 134, 135, 136, 164, and 171) achieved an AUC of 0.6 or higher (predicting active compounds at random would be expected to yield an AUC of 0.5), and only six (IDs 135, 136, 147, 148, 164, and 172) achieved an enrichment factor at 10% (EF) more than one standard error better than random (1.0). Enrichment plots for two of the top submissions (164, which is top by every metric, and 135,

[2] The challenge began with 322 compounds, 260 non-binders, and 62 binders, but due to errors and redundancies, final analysis was run on 305 compounds and 56 binders.

ID	BEDROC	AUC	EF (10%)
007	0.20+/-0.07	0.53+/-0.04	1.07+/-0.37
008	0.37+/-0.09	0.63+/-0.04	1.97+/-0.48
133	0.13+/-0.05	0.59+/-0.04	1.07+/-0.39
134	0.20+/-0.07	0.62+/-0.04	1.25+/-0.40
135	0.37+/-0.09	0.64+/-0.04	2.21+/-0.51
136	0.27+/-0.08	0.65+/-0.04	1.47+/-0.43
146	0.20+/-0.07	0.56+/-0.04	1.25+/-0.39
147	0.28+/-0.08	0.55+/-0.05	1.79+/-0.46
148	0.36+/-0.09	0.59+/-0.04	1.61+/-0.45
157	0.14+/-0.07	0.45+/-0.04	0.72+/-0.32
164	0.56+/-0.09	0.71+/-0.04	3.22+/-0.49
165	0.20+/-0.06	0.55+/-0.04	1.43+/-0.44
171	0.38+/-0.10	0.60+/-0.04	1.25+/-0.41
172	0.24+/-0.07	0.54+/-0.04	1.43+/-0.42
173	0.18+/-0.06	0.49+/-0.05	1.25+/-0.41
174	0.20+/-0.07	0.49+/-0.05	1.07+/-0.39
175	0.11+/-0.05	0.45+/-0.04	0.72+/-0.32
176	0.10+/-0.05	0.46+/-0.05	0.36+/-0.26
198	0.13+/-0.06	0.52+/-0.04	0.72+/-0.34
200	0.14+/-0.07	0.48+/-0.04	0.54+/-0.29
238	0.14+/-0.07	0.50+/-0.04	0.72+/-0.34
239	0.19+/-0.07	0.56+/-0.04	1.07+/-0.38
240	0.20+/-0.07	0.53+/-0.04	1.25+/-0.41
241	0.16+/-0.06	0.54+/-0.04	0.89+/-0.36
242	0.16+/-0.06	0.56+/-0.04	0.89+/-0.38
524	0.11+/-0.05	0.48+/-0.04	0.72+/-0.34
546	0.20+/-0.07	0.56+/-0.04	1.25+/-0.41
547	0.20+/-0.07	0.57+/-0.04	1.25+/-0.39

Table 2.1: Calculated metrics for SAMPL4 virtual screening submissions. Also shown are control or null models 007 and 008. For each submission ID, we computed the area under the enrichment curve (AUC), the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC), and the enrichment factor at 10%. For this set, the maximum enrichment factor at 10% is 305/56 = 5.45.

which is consistently among the top) are shown in Figure 2.3, along with an enrichment plot for a more typical submission (198).

Submission 164 was the top performer by every metric, and really stood out from the pack, especially in terms of early enrichment, so it is worth examining the approach in slightly more detail, but we refer the reader elsewhere for a full description[174]. In brief, this submission came from a human expert with more than 10 years experience working on this specific target. The specific procedure used docking with GOLD, then a pharmacophore search done in MOE using many crystallographic structures of LEDGF ligands to generate the query. MOE and an electrostatic similarity search were used for filtering. The correct stereochemistry of binders was assigned manually after electrostatic similarity comparison and binding mode examination. Overall, screening via this approach involved substantial manual intervention and expert knowledge. It is worth highlighting that this approach did especially well at early enrichment, with an enrichment factor of 3.2+/-0.5. The maximum EF at 10% on this set is 5.2. The observation that the top performing submission used substantial manual intervention and human expertise echoes the conclusion of SAMPL2, where human experts outperformed automated methods at pose prediction[159].

Submission 135 was also particularly interesting, in that it began from essentially the same inputs as 133 and 134 – AutoDock/Vina docking calculations – but used BEDAM alchemical binding free energy calculations[50, 23] to score predictions. This appears to have been remarkably successful at improving recognition of LEDGF binders, and was hampered by time constraints – not all molecules could be analyzed in this way, so apparently many of the actives which were still missed lacked binding free energy estimates. We refer the reader elsewhere in this issue for additional discussion of this submission[48].

Overall, we saw submissions using a fairly wide range of other methods, though in general most of these were relatively rapid methods (with the exception of 135) involving at least some component of docking. A variety of submissions used simple docking with various packages and different target protein structures (133, 157, 198, 200, 238, 239-242, 524, 546-547) and most others used docking plus something else (i.e. rescoring, scoring function modifications, etc.). For example, as discussed, 135 used docking plus alchemical free energy calculations, while 136 used a consensus score of 133-135, 146-148 used WILMA docking plus SIE re-scoring, 165 used protein-specific charges, and so on. 172-176 stood out from other approaches because they used a pharmacophore docking approach. However, in general among these methods, we do not see an approach which clearly stands out from the rest.

We also ran two control or null models, submissions 007 and 008, which were not formally SAMPL submissions. ID 007 is based on molecular weight alone – compounds are ranked simply based on molecular weight, with heavier compounds predicted to bind best. ID 008 is based on ligand shape similarity, computing using OpenEye's ROCS, with reference ligand CDQ 225 from the 3NF8 reference structure. This approach, shape similarity to a known ligand, is actually quite reasonable and should be thought of as a control rather than a null model. Indeed, we find that many methods outperform 007, which does not do significantly better than random at recognizing actives. On the other hand, 008, based on shape similarity to a known ligand, performs quite well, and indeed is among the top methods in terms of early enrichment and is one of the approaches achieving an AUC over 0.6.

2.5.2 Post-SAMPL: Alternate isomers may not be non-binders

We constructed the virtual screening set with the assumption that alternate isomers of binders are in fact non-binders, but this may in fact be an oversimplification. This approach seemed reasonable initially, since SPR and crystallography were typically run on mixtures of isomers, so isomers which were not observed to bind crystallographically are at least much weaker binders than the binding isomer. But this does not guarantee that they are actually non-binders. Consider a hypothetical molecule A with isomers A_0 and A_1, where A_0 has a dissociation constant of $5\mu M$ and A_1 has a dissociation constant of $100\mu M$. Binding of A_1 is sufficiently weaker than A_0 that it would be extremely difficult to detect in an assay on an equal mixture of the two isomers, and hence would be labeled a "non-binder". Hence, it would perhaps be more appropriate to divide our virtual screening set into three categories: "actives", "inactives", and "inactives or very weak actives". Success in the last category would require a method to rank these compounds lower than the corresponding alternate isomers which are in the "actives" category. In any case, this analysis suggests that a re-analysis of the SAMPL results may be needed.

In view of this uncertainty, we ran a re-analysis of the virtual screening challenge on a new set which dropped all alternate isomers of active compounds, effectively excluding the "inactive or very weak active" category and retaining only true actives and inactives. This reduced the number of compounds analyzed from 305 to 189, while retaining the same 56 actives. Full statistics and plots for this subset are provided in the Supporting Information. Overall, the ranking of methods by our different metrics stayed somewhat similar in many cases, though the best BEDROC values rose very substantially, indicating better early enrichment. Also, submission 547, which had been essentially in the middle by every metric instead jumped to second place by every metric, in some cases within error of our best submission, 164. In our view the marked change in performance here suggests that some fraction of the inactive compounds may in fact be weak actives. Additionally, this observation has obvious implications for future experimental design and design of SAMPL challenges, since it means additional information is needed to distinguish between very weak actives which are tested together with stronger actives, versus true non binders.

2.6 Integrase Pose Prediction Results

Pose prediction participation was, from our perspective, surprisingly light. We received 12 submissions from five research groups. While in principle participants could submit multiple predicted binding modes for each ligand (since some ligands bound in multiple sites, completely successful predictions would have needed to do so), only three submissions did so, and in only a few cases. As noted above, we score each method both by the best predicted pose for each ligand, and by the best predicted pose for each experimental binding mode. Since a number of ligands have multiple binding modes, the latter is a substantially longer set.

Our initial analysis focused primarily on examining the root mean squared deviation (RMSD) for each submission, as shown in Table 2.2 and Figure 2.4. Because RMSD is unbounded, a simple mean is not necessarily a good metric overall, so we also looked at the median RMSD. As discussed in the analysis section above, we also wanted to look at how often binding modes were predicted successfully, but without an arbitrary "success" RMSD threshold. So instead, we computed the area under the curve (AUC) for the fraction of poses predicted correctly at a given cutoff level; here, a higher number is better.

As Figure 2.4 shows, each method had substantial variability in performance. While the top methods tended to predict more poses correctly, no method predicted all binding modes to high accuracy, as seen by RMSD. The figure focuses on the best predictions for each ligand, but a similar conclusion holds for predictions when judged by pose, as shown in the Supporting Information. Still, by a variety of metrics, submissions 177, 536, 143, and 154 were typically among the top performers. Test submission 301 also did quite well, and is a reference model we ran internally and will be discussed below. ID 177 used XP Glide[46] with rescoring via DrugScore and MM-PB/SA[96], and 143 used AutoDock Vina[167], while ID 154 used Wilma docking and SIE re-scoring[128, 164]. ID 536 used DOCK 3.7. All of these except submission 536 considered binding to multiple sites.

Our AUC metric assesses what fraction of pose predictions were successful as a function of the definition of success (RMSD cutoff). RMSD has at least one major disadvantage, in that it is unbounded, so a method which performs modestly well on 25 compounds and very poorly on 30 could actually appear worse (by average RMSD) than a method which performs fairly poorly on all 55, simply because very large RMSD values can contribute so much to the average. In contrast, AUC is relatively insensitive to failures and particularly sensitive to the fraction of binding poses correctly identified. Sample plots of the data which goes into the AUC calculation are shown in Figure 2.5. The plot uses a semi-log scale, and clearly shows how submission 177 performs substantially better than 583 in terms of pose prediction success.

Our fingerprint Tanimoto metric focuses on whether poses identified the correct interactions and contacts with the protein, rather than on reproducing the experimental binding mode precisely. Thus, this is a more flexible criteria for success within the binding site, though it rapidly goes to zero as the predicted binding mode moves away from the true binding site. The top methods seem to have substantial success, typically, at identifying the correct interactions, and a number of submissions perform nearly as well by this method and are probably statistically indistinguishable.

Some insight into why methods had such a broad range of performance can be gained by examining Figure 2.4(c), which looks only at the fraction of ligands placed into the correct binding site. Since most methods considered all three binding sites, many of the high RMSD predictions were a result of predicting the wrong binding site. Thus, performance seems more comparable across methods when considering only poses within the correct site. However, it is worth noting that the same submissions are still among the top performers. This is further highlighted by looking at the number of ligands placed into the correct site, in Figure 2.4(f). Despite the hint given participants that they could focus primarily on the LEDGF site, most chose to include both the Y3 and FBP sites when making predictions, as well, and so the majority of submissions typically selected the incorrect binding site. However, of the top submissions, most *did* include multiple sites in their analysis. The fact that several submissions were thus fairly successful at identifying the correct binding site is encouraging.

Submissions 300 and 301 are an attempt at generating null or comparison models. Thus, these are not discussed above because these are test cases not formally submitted for SAMPL, but some discussion is warranted. Both of these submissions were done in a blind manner, just as the rest of the SAMPL, and submission 300 (but not 301) was done prior to the submission deadline. Submission 300 was a control run in which a beginning high school student in the Mobley laboratory predicted binding modes using AutoDock simply by following online tutorials and documentation, with no separate instruction and with minimal background reading on IN and on this particular series. It appears that one major challenge for 300 was the definition of and identification of the binding site region. Very few ligands were placed nearest the correct site, and even of those, none came close to the correct binding modes. Primarily, this probably serves to illustrate that some expertise in docking and some knowledge of likely binding sites is still needed for successful pose prediction.

ID 301 provides a more challenging benchmark. This applied a different approach than most participants, and took all six bound ligands out of the 3NF8 reference structure. Each ligand was then shape-overlaid onto the 3NF8 ligands using OpenEye's ROCS, and bumping poses were removed. Each pose was then energy minimized with MMFF, and the remaining pose with the best MMFF energy was then scored. This actually would have ended up being the top submission by most metrics, and does extremely well. This is partly because this is precisely the type of challenge where a ligand-based approach such as this one ought to do well – where there are structures of related ligands bound in all the binding sites of interest – and partly because the LEDGF site seems to have typically resulted in the best MMFF energy.

ID	$\overline{RMSD}$	med. $RMSD$	AUC	$\overline{RMSD}$	med. $RMSD$	AUC	$\overline{RMSD}$	med. $RMSD$	AUC
	by ligand			by pose			correct site, by ligand		
143	6.5+/-1.0	3.8+/-0.4	93.4+/-1.0	7.2+/-0.8	4.1+/-0.4	92.7+/-0.8	3.4+/-1.0	3.4+/-0.4	96.5+/-1.0
154	12.2+/-1.7	4.4+/-4.7	87.7+/-1.7	13.0+/-1.2	4.9+/-4.7	86.9+/-1.2	2.3+/-1.7	1.5+/-4.7	97.6+/-1.7
155	15.4+/-1.5	17.2+/-5.5	84.5+/-1.5	15.8+/-1.1	20.0+/-4.1	84.1+/-1.1	3.8+/-1.5	1.8+/-5.5	96.1+/-1.5
156	7.8+/-0.9	5.8+/-1.0	92.1+/-0.9	8.6+/-0.7	6.2+/-0.8	91.3+/-0.7	6.7+/-0.9	5.2+/-1.0	93.2+/-0.9
177	5.6+/-0.9	4.0+/-0.4	94.3+/-0.9	6.4+/-0.7	4.1+/-0.2	93.5+/-0.7	3.3+/-0.9	3.7+/-0.4	96.6+/-0.9
300	20.4+/-1.1	18.7+/-0.4	79.5+/-1.1	20.4+/-0.7	18.9+/-0.3	79.5+/-0.7	23.2+/-1.1	17.5+/-0.4	76.7+/-1.1
301	4.3+/-0.8	2.8+/-0.4	95.6+/-0.8	5.3+/-0.7	2.8+/-0.3	94.6+/-0.7	2.8+/-0.8	2.5+/-0.4	97.1+/-0.8
535	22.6+/-0.8	24.0+/-0.3	77.3+/-0.8	22.7+/-0.5	24.0+/-0.2	77.2+/-0.5	3.4+/-0.8	3.3+/-0.3	96.5+/-0.8
536	6.3+/-0.7	4.8+/-0.4	93.6+/-0.7	7.2+/-0.6	4.9+/-0.3	92.7+/-0.6	5.0+/-0.7	4.7+/-0.4	94.9+/-0.7
537	26.7+/-0.9	28.8+/-0.6	73.2+/-0.9	26.9+/-0.6	28.6+/-0.4	73.0+/-0.6	6.3+/-0.9	6.3+/-0.6	93.5+/-0.9
538	26.1+/-1.1	28.1+/-0.5	73.8+/-1.1	26.3+/-0.7	28.1+/-0.3	73.6+/-0.7	5.2+/-1.1	5.1+/-0.5	94.7+/-1.1
539	27.3+/-1.0	29.2+/-0.5	72.6+/-1.0	27.4+/-0.6	29.0+/-0.3	72.5+/-0.6	5.5+/-1.0	6.2+/-0.5	94.4+/-1.0
540	27.0+/-0.9	29.1+/-0.4	72.9+/-0.9	27.2+/-0.6	29.0+/-0.3	72.7+/-0.6	5.5+/-0.9	6.2+/-0.4	94.4+/-0.9
583	25.7+/-1.6	24.3+/-2.7	74.2+/-1.6	26.0+/-1.1	25.3+/-1.8	73.9+/-1.1	13.5+/-1.6	14.1+/-2.7	86.4+/-1.6
1000	20.7+/-1.4	24.4+/-0.8	79.2+/-1.4	20.9+/-0.9	24.4+/-0.4	79.0+/-0.9	4.0+/-1.4	4.5+/-0.8	95.9+/-1.4

Table 2.2: Statistics for SAMPL4 pose prediction. Shown are statistics by ligand (where the lowest RMSD prediction is taken for each ligand), by pose (where the lowest RMSD prediction is considered separately for each experimental binding mode), and by ligand for only the fraction of ligands placed into (or nearest) the correct binding site.

We examined median error across different ligands in the set to try and understand whether particular classes of ligand were especially difficult to predict. However, almost every ligand is well predicted by at least some methods. Median errors across all methods do fluctuate substantially from ligand-to-ligand, but we did not immediately observe patterns where particular classes or groups of ligands were particularly difficult to predict. We did observe a slight correlation between increasing molecular weight and median RMSD, but some correlation between RMSD and molecular weight is to be expected regardless. We also find (as did Coleman et al.[27]) a slight trend that more highly charged ligands (charge -2, or zwitterions with charge -2 + 1) may have higher median RMSDs, but the test set is small enough it is hard to be sure this is statistically significant.

2.7 Integrase Affinity Results

The integrase affinity challenge received 15 submissions from four groups. Statistics are shown in Table 2.3 and Figure 2.6. We used the Kendall W statistic to see whether there was a clear leader and arrived at a value of $W = 0.80 \pm 0.08$, indicating that almost all affinities are better predicted by one submission than any other. This submission was 184 (Figure 2.7), which used the SIE scoring function[128, 164] with the FiSH[28] hydration model, but even this suffered from rather poor performance overall. While this submission's RMS error, 1.83 ± 0.41 kcal/mol, and the AUE 1.33 ± 0.44 kcal/mol seem acceptable, the experimental data spans only a 1 kcal/mol range, so the error is larger than the signal. Thus, for ID 184, the Pearson R and Kendall tau are actually negative, indicating incorrect ranking. Interestingly, one group of compounds seems well predicted, while the rest are very poorly predicted. The authors suggest that some of this noise can be reduced by using a common protein structure instead of the cognate crystallographic structures[71].

In this challenge component, most submissions actually used docking to try and predict affinities. And submission IDs 199, 201, and 233-237 actually submitted scores from the DOCK package as "affinity" predictions. Since these scores are not normalized, the hope was that these would provide some correlation with experiment, rather than actually provide reasonable affinity estimates, hence the very large errors for these submissions. One notable exception to the typical docking approach here was submissions 190-191, which used an MM-PB/SA approach.

Submission IDs 012 and 013 were a null model based on the classic work of Kuntz[99], where affinities were predicted based on the number of heavy atoms with a value of 1.5 kcal/mol per heavy atom up to a plateau value, and then were a constant beyond that (ID 012). Because all of these ligands are large enough to have reached the plateau, this resulted in a constant prediction for all ligands, so model 013 removes the plateau. These were provided by Coleman and collaborators[27]. It is worth noting that a variety of other methods substantially outperform these null models here, despite the limited nature of this test.

Overall, given the very narrow range of experimental binding free energies for these few

ID	Avg. Err.	RMS	AUE	tau	R
012	-18.0+/-0.1	18.0+/-0.1	18.0+/-0.1	-0.2+/-0.0	-0.2+/-0.0
013	-41.6+/-2.5	42.2+/-2.8	41.6+/-2.5	-0.4+/-0.0	-0.6+/-0.0
182	-3.8+/-0.3	3.9+/-0.3	3.8+/-0.3	-0.4+/-0.3	-0.7+/-0.3
183	-3.2+/-0.3	3.3+/-0.3	3.3+/-0.3	-0.6+/-0.2	-0.8+/-0.2
184	-1.2+/-0.5	1.8+/-0.4	1.3+/-0.4	-0.1+/-0.3	-0.3+/-0.3
190	-2.2+/-0.8	3.1+/-0.5	2.7+/-0.5	-0.1+/-0.2	-0.2+/-0.3
191	-3.8+/-0.3	4.0+/-0.3	3.8+/-0.3	0.2+/-0.3	0.4+/-0.3
199	-22.2+/-4.2	25.1+/-2.8	23.0+/-3.6	0.6+/-0.3	0.6+/-0.4
201	-25.1+/-5.0	28.8+/-2.8	27.1+/-3.4	0.5+/-0.3	0.6+/-0.4
233	-24.6+/-4.8	28.1+/-2.6	26.6+/-3.3	0.4+/-0.3	0.5+/-0.4
234	-25.6+/-4.4	28.5+/-2.8	26.5+/-3.7	0.2+/-0.3	0.5+/-0.4
235	-17.4+/-3.6	20.1+/-2.5	18.2+/-3.1	0.5+/-0.3	0.6+/-0.3
236	-17.5+/-3.4	20.0+/-1.9	18.9+/-2.4	0.4+/-0.3	0.6+/-0.4
237	-27.0+/-3.4	28.6+/-2.7	27.0+/-3.4	0.2+/-0.3	0.5+/-0.4
549	-7.8+/-1.3	8.6+/-1.3	7.8+/-1.3	-0.2+/-0.3	-0.3+/-0.3

Table 2.3: Statistics for IN affinity prediction. Shown are the average error, RMS error, AUE, Kendall tau, and Pearson R.

relatively weak ligands, it is difficult to draw any strong conclusions from this portion of the challenge.

2.8 Conclusions

Overall, the HIV integrase portion of the SAMPL4 challenge proved extremely challenging. The virtual screening component was difficult apparently because the inactive compounds are true inactives and were so similar to active compounds, and indeed were tested precisely because they were thought to bind. Thus, it proved extremely difficult to substantially enrich compounds in this portion of the challenge. Likewise, the binding mode prediction of the challenge was difficult, partly because of the several binding sites participants had to deal with. And the narrow range of relatively weak affinities in the affinity prediction challenge made it challenging to achieve any correlation between calculated and experimental values – though several methods did have reasonably low errors.

However, it was encouraging that some methods were able to significantly enrich actives in the virtual screening portion of the challenge. Here, one method in particular stood out from the pack, and interestingly, it involved substantial manual intervention from a human expert in the screening process. Apparently, human expertise still pays off. To us, this is actually somewhat encouraging, in that it means that there is still more we can teach our binding prediction algorithms.

For binding mode prediction, all methods performed poorly on at least some ligands, but one major source of large errors was placement of ligands into the incorrect binding site (since three binding sites were possible). Interestingly, however, placing ligands into the correct binding site was not a guarantee of success, and some of the top methods actually considered binding to all three sites. This suggests that, at least in some cases, binding site identification may be possible with today's methods. Interestingly, a control model we ran using a ligand-based approach actually performed quite well at this portion of the challenge, suggesting that in the future, participants may want to consider alternate approaches such as this to help their structure-based efforts. A similar (control) ligand-based approach also performed well in the virtual screening test, further supporting this line of thinking. Possibly in future challenges the best approach may involve a combination of methods.

Overall, we believe the integrase component of the SAMPL4 challenge was a valuable test, and we are convinced that blind tests like this are a helpful way to gain insight into how methods may perform in a real-world discovery setting. Expert knowledge does seem to continue to play an important role, but it does not guarantee success, nor does the lack of expert knowledge guarantee failure. Much depends on both the practitioner and the details of the approach.

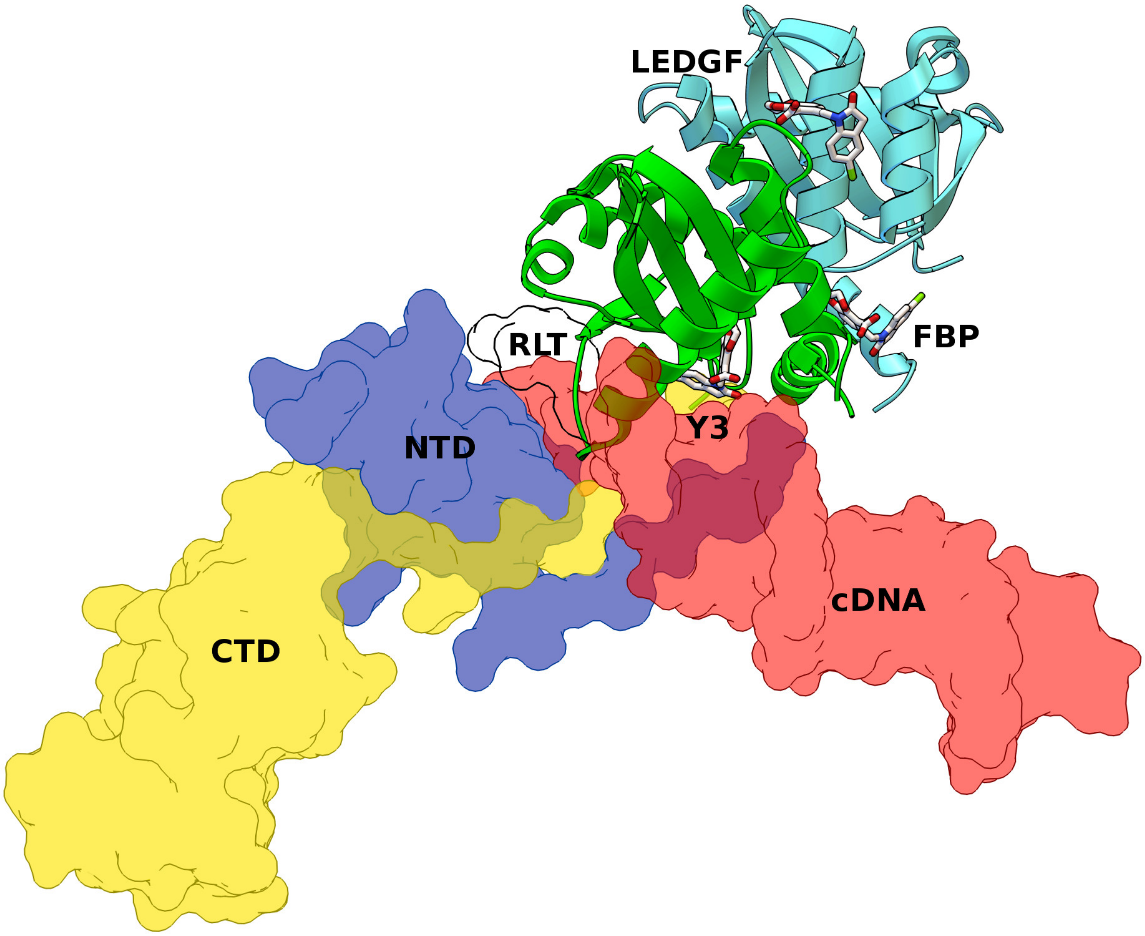
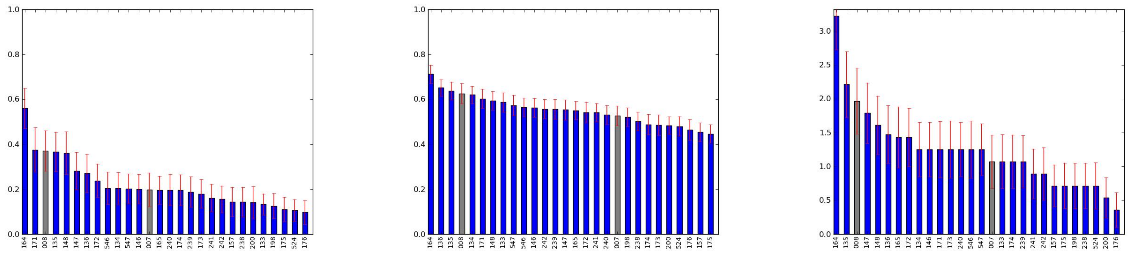


Figure 2.1: Integrase (IN) functional structure and architecture. The three domain structure of a monomer of IN is displayed. The PFV IN crystal structure (from 3OS1.pdb) of the "target capture complex" with DNA is displayed in surface mode, with the C-Terminal Domain in yellow, the N-Terminal Domain in light blue, and the human DNA in salmon. The 3NF8 reference structure of the HIV IN Catalytic Core Domain dimer was superimposed onto this PFV IN crystal structure, and its CCD is shown in ribbon mode (with one monomer in green and the other in cyan). The "CDQ" allosteric fragment from 3NF8 is displayed as sticks with white carbons to highlight the three allosteric sites of HIV IN that were part of the SAMPL4 challenge: LEDGF, Y3, and FBP. A black outline and the label RLT show the location of the active site of IN. Raltegravir (labeled as RLT) was extracted by superimposing the PFV IN crystal structure from 3OYA.pdb onto the 3OS1.pdb structure of PFV IN. During catalysis, HIV IN is present as a tetramer (i.e., a dimer of dimers).

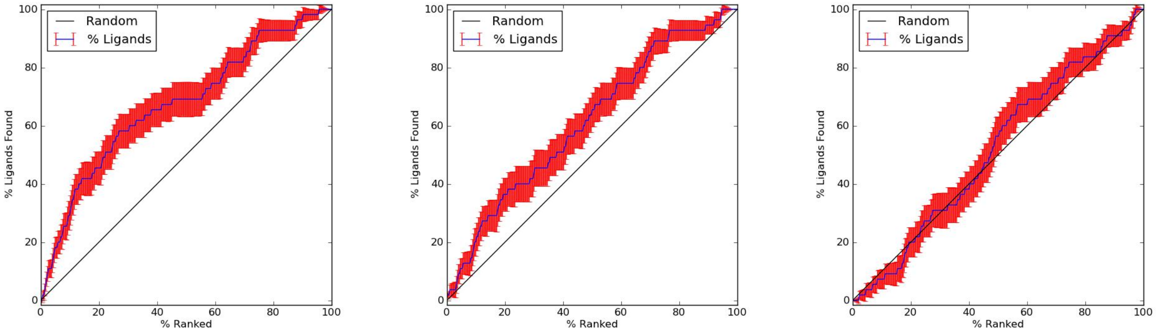


(a) BEDROC

(b) AUC

(c) EF 10%

Figure 2.2: Calculated metrics for SAMPL4 virtual screening statistics, graphed in ranked order. The statistics are as given in Table 2.1. Note that many submissions have overlapping error bars, so ranked order is not necessarily indicative of *significantly* better performance.

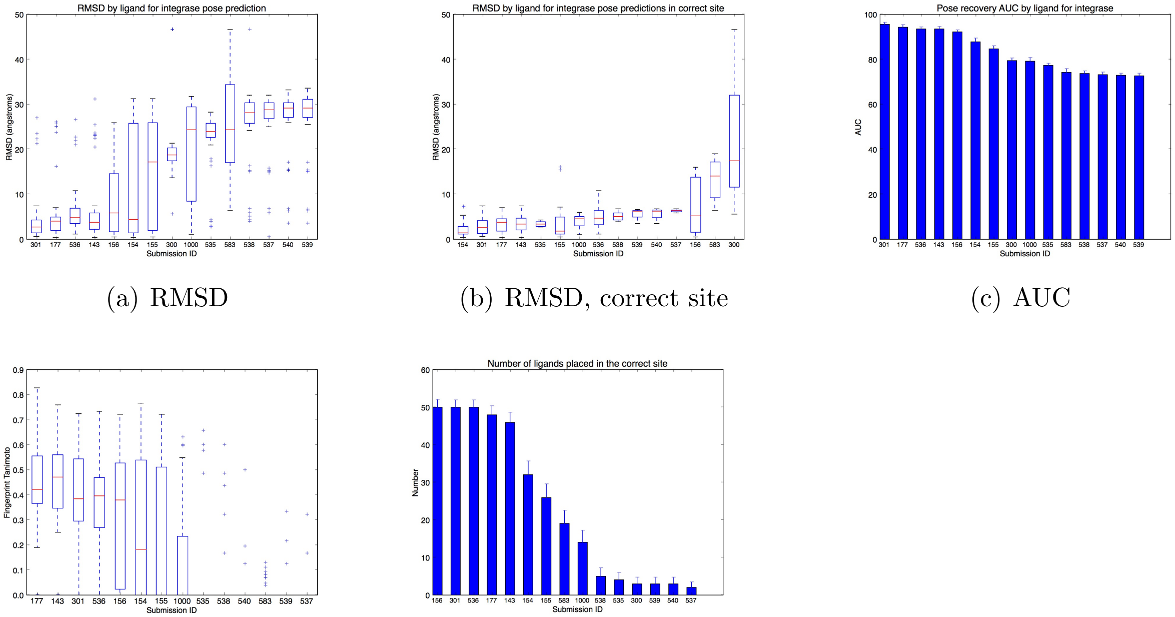


(a) Enrichment, 164

(b) Enrichment, 135

(c) Enrichment, 198

Figure 2.3: Enrichment plots for two of the best-performing virtual screening submissions, and one for which performance is close to random. Submission 164 was the top performer by all metrics, and 135 was one of the other top performers. 198 is shown here as representative of more typical performance for comparison. Error bars are shown in red and give an idea of the expected variation with the composition of the set.



(a) RMSD

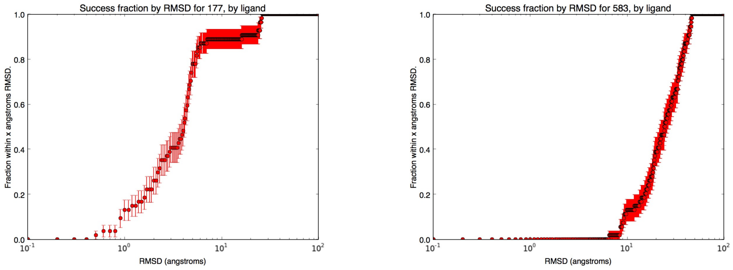
(b) RMSD, correct site

(c) AUC

(d) Fingerprint Tanimoto

(e) Site identification

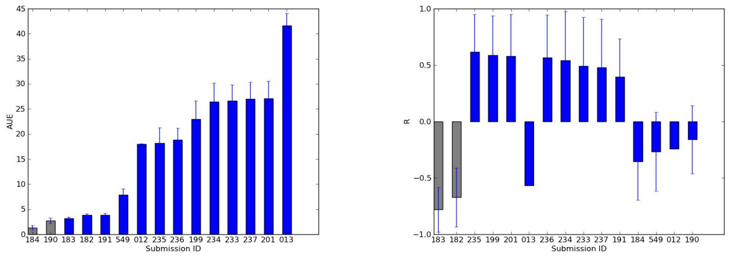
Figure 2.4: Ranked performance on pose prediction, by various metrics. (a)-(b), box/whisker plots showing performance by ligand as judged by RMSD; (b) focuses only on the subset of ligands placed within the correct site. (c) shows performance judged by AUC, by ligand; and (d) shows performance by ligand as judged by interaction fingerprint Tanimoto scores. (e) shows the number of ligands placed into the correct binding site for each submission. For bar plots, normal submissions are shown in blue, while control models (300, 301) are shown in gray, as discussed in the text.



(a) 177

(b) 583

Figure 2.5: Fraction of ligands with correct binding modes, versus RMSD cutoff. Shown are the fraction of ligand binding modes predicted correctly within a cutoff of x Å RMSD, where x is the horizontal axis. The scale is semi-log. Method 177 performed particularly well by this and other metrics, while 583 did not perform particularly well here. In submission 177, most ligands are predicted correctly within 10 Å RMSD, with a substantial fraction better 3 Å. In contrast, in 583, only a small number are predicted better than 10 Å RMSD.



(a) Ranked AUE

(b) Ranked R

Figure 2.6: Representative statistics for the integrase affinity challenge. Shown are the Pearson correlation coefficient, R, and the average unsigned error (AUE). Normal submissions are shown in blue; control models are shown in gray, as discussed in the text.

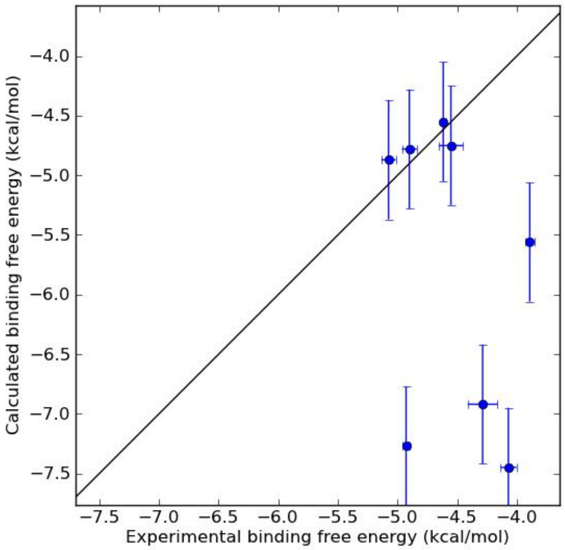


Figure 2.7: Performance of submission 184 in the integrase affinity challenge. While one group of compounds was well predicted, another group was not. 184 was essentially the top submission in the challenge.

Chapter 3

Lead Optimization Mapper: Automating free energy calculations for lead optimization

3.1 Abstract

Alchemical free energy calculations hold increasing promise as an aid to drug discovery efforts. However, applications of these techniques in discovery projects have been relatively few, partly because of the difficulty of planning and setting up calculations. Here, we introduce Lead Optimization Mapper, LOMAP, an automated algorithm to plan efficient relative free energy calculations between potential ligands within a substantial library of perhaps hundreds of compounds. In this approach, ligands are first grouped by structural similarity primarily based on the size of a (loosely defined) maximal common substructure, and then calculations are planned within and between sets of structurally related compounds. An emphasis is placed on ensuring that relative free energies can be obtained between any

pair of compounds without combining the results of too many different relative free energy calculations (to avoid accumulation of error) and by providing some redundancy to allow for the possibility of error and consistency checking and provide some insight into when results can be expected to be unreliable. The algorithm is discussed in detail and a Python implementation, based on both Schrödinger's and OpenEye's APIs, has been made available freely under the BSD license.

3.2 Introduction

A good deal of early-stage drug discovery focuses on finding small molecules with suitable affinity for a target binding site in a receptor, while simultaneously having good physical and pharmacological properties to make orally available drugs[41, 147, 107]. The very earliest stages of the process involve finding hits, small molecules which bind to the target receptor relatively weakly. Then, these hits need to be turned into leads – molecules which have suitable properties to potentially become drugs while also having sufficient affinity for the target receptor [41, 107].

Ideally, computational methods could play a role in guiding early stage drug discovery, which has traditionally been slow and time-consuming, filled with trial and error. Even the process of finding molecules which bind with sufficient affinity to the target compound can be slow and involve synthesizing hundreds of molecules [147, 183] resulting in substantial costs, both in terms of material and in time spent[41, 161]. But developing computational tools with sufficient accuracy to reliably guide this discovery and optimization process has proven challenging [23, 183]. Some of the most popular methods, such as docking, while seeing widespread use, do not reliably yield any correlation between predicted binding strength and experimental data [183]. And the path to improve these methods has often been unclear, in part because of the number of approximations made. So, while computational methods are used in early stage drug discovery, in most discovery projects they do not play a key role in *guiding* the process [40, 183, 139, 144]. More quantitative methods could have a more dramatic impact on the early stage discovery process.

Some methods promise higher accuracy, and thus the potential for more of a role in guiding early stage drug discovery. More rigorous methods for binding affinity prediction are available, such as alchemical binding free energy techniques. These compute binding free energies, or differences in binding free energies, from molecular simulations using techniques based on perturbations [24, 116, 54, 152]. These techniques can be used to compute both absolute binding free energies (ABFE) [14, 124, 34] or, more commonly, relative binding free energies (RBFE) between related inhibitors [161, 162, 24, 116, 23]. While in a number of cases these techniques show considerable promise, a key obstacle hampering their more widespread use has been the difficulty of setting up and performing these calculations, which typically requires considerable expert intervention [21, 22]. So, while more rigorous methods are available, they, too, do not typically help guide drug discovery[23].

Here, our focus is on automating setup of alchemical relative free energy calculations, allowing their application to large numbers of molecules with relatively little human intervention in a discovery-type setting. We had previously worked only on small numbers of molecules, and the overhead involved in setup was not a huge concern. However, an abrupt collision with the realities faced every day by modelers in the pharmaceutical industry helped us realize we need to do better, as we sought to use free energy calculations to screen a modest library of potential inhibitors. Essentially, our problem (not unlike that facing many early stage drug discovery projects) was this: Given a set of knowns, and a library of tens to a few hundred potential other molecules of interest (some related to knowns, some not related), predict which of this library might be best to follow up on experimentally. The relevance of this task to drug discovery motivated the present development of an automated setup tool for alchemical free energy calculations. The importance of this problem has also similarly motivated a recent tool for automated setup of endpoint free energy calculations[72].

In planning calculations of binding free energies for these compounds, we chose to compute RBFE, comparing binding strengths of related compounds, rather than computing ABFE for individual compounds. Several reasons motivated this choice. First, RBFE calculations between related compounds are often considered more efficient than ABFE calculations since they involve insertion and deletion of relatively few atoms, historically one of the most computationally expensive steps [5, 188, 163]. Also, any protein motions or conformational changes that happen on binding but are common to all ligands do not necessarily need to be sampled. Second, many molecules in our initial set were charged, and free energy calculations involving changes in the system net charge, as we would be doing in ABFE calculations, pose technical challenges that are not yet well understood for systems more complicated than individual ions in water [87, 88, 76, 77]. Preserving the net charge of the system by doing relative free energy calculations between molecules sharing the same net charge bypasses these problems.

It is worth noting that RBFE calculations do have one major limitation, in that they do require knowledge of the likely binding mode of the compounds of interest. Such knowledge will often be available in structure-based drug design projects, especially at the lead optimization phase, but it is worth noting this requirement. This challenge is not unique to RBFE calculations, though – the same issue also confronts most other techniques, including ABFE calculations, which also must either take the likely binding mode as input, or at least sample it adequately in the resulting simulations. However, in cases where there is substantial uncertainty as to the compound's likely binding mode and multiple possibilities are available, ABFE calculations may actually be preferable to RBFE calculations as multiple binding modes can be difficult to handle within the RBFE framework [125]. Overall, though, it is generally thought that RBFE calculations are easier and more efficient than ABFE calculations.

While RBFE calculations avoid some problems with ABFE calculations, they do require a planning step that is not needed for ABFE calculations: For a possible 50 compound lead series, which of $50 \times 49/2 = 1225$ possible relative free energy calculations should we actually do? Each relative free energy calculation compares binding of a pair of inhibitors, so we need an automated way to decide which pairs of inhibitors we ought to plan relative calculations between. Rather than doing 1225 RBFE calculations, we ought to be able to span the entire library with just over 50 relative calculations, yielding relative free energies for all of the molecules. Hence, our main focus here is development of a tool which can automatically plan RBFE calculations spanning a library of compounds.

3.3 Method

3.3.1 Design goals

What criteria make for a good relative free energy calculation? That is, for a library of n compounds, with $n \times (n-1)/2$ possible relative free energy calculations, which calculations can we expect to actually work reasonably well? Some choices are clearly bad. For example, if two molecules share absolutely no atoms in common, computing an RBFE between the two requires subtraction of two ABFE calculations. Beyond this obvious consideration, we specified a number of other design goals. These are broadly based on efficiency considerations identified in the literature, and also focus to some extent on minimizing error accumulation and building in ways to identify calculations which may be wrong due to poor convergence.

Goal 1: Compounds being compared should be as similar as possible, minimizing atomic deletions and insertions

Generally, we would prefer the compounds being compared to be as chemically similar as possible, maximizing the likelihood that they have the same or similar binding modes and minimizing the number of atomic deletions and insertions required [5, 188, 163, 8, 81]. For example, changes in atomic partial charge and atom type can be typically done using molecular dynamics simulations involving relatively few intermediates (λ values) spanning between the two end states, perhaps even as few as 5-11[103, 116, 104, 180]. Insertions of individual atoms can also be done in a relatively straightforward way[81, 163, 8] but deletions or insertions of entire functional groups may require as much as 2-3 times as many simulations[155, 156, 122, 5, 188, 163] because of the need to spread out deletions/insertions across multiple simulations. Additionally, these larger chemical changes typically lead to larger free energies, and hence the chance of larger errors. So, a major goal is to plan relative calculations between the most similar molecules, minimizing the number of atomic deletions and insertions.

It is worth noting, however, that from the standpoint of classical fixed-charge simulations, "similar" means something slightly different than it typically does in drug discovery. Specifically, in simulations, atom type changes are quite straightforward, and need not be avoided. For example, changing benzene into chlorobenzene is easy – no atoms are deleted or inserted, and there is only a modification of the atomic partial charges and minor changes to some Lennard-Jones parameters. Similarly, changing a nitrogen-containing heterocycle into a sulfur-containing one is straightforward as long as the pattern of connected atoms is the same. So we are willing to grant considerable chemical leeway when deciding which molecules are "similar". Figure 4.1 shows an example of a variety of favorable transformations beginning from 2-methylnaphthalene.

Goal 2: Rings should be preserved as much as possible

The larger the functional group being deleted, the larger the potential problems with insertion/deletion, so we prefer to avoid deletion and insertion of ring systems as much as possible (though this is to be preferred over breaking or forming rings[153], as noted below). Additionally, deletion of large bulky functional groups such as rings may provide a molecule with substantially more room in a binding site, reducing the likelihood that it will remain in the expected binding mode, and increasing the potential for problems adequately sampling potential binding modes [125]. Hence, we choose to try and retain ring systems as much as possible.

Goal 3: Ligands being compared must share the same net charge

As noted above, for technical reasons, relative free energy calculations involving changes of the net charge of a system are to be expected to be unreliable. Specifically, changing one charged ligand into a ligand of a different charge in the binding site leads to a contribution to the free energy due to the change in charge and its interaction with the surrounding solvent dielectric, periodic copies of the system, and other factors [87, 88, 76, 77]. This would cancel out if these contributions were the same in solvent, but in general they are not, partly due to differences in system size and composition in the two environments. Therefore, we choose to plan relative calculations only between subsets of ligands of the same net charge. To allow estimation of the affinity of all compounds, selected compounds of known affinity could be included in each subset, if such compounds are available (see Section 3.3.2). In our view, RBFE calculations for chemical modifications which change the net charge (such as addition of a carboxylic acid) must wait for algorithmic developments.

Goal 4: Portions of multi-ring systems can only be deleted if rings are planar, and this should be avoided when possible

Alchemical free energy calculations rely on a thermodynamic cycle, which must properly close to yield relative free energies. When deleting atoms, we leave behind so-called “dummy atoms” which still have bonds to the remainder of the molecule but do not interact with the system in any other way[153]. For our thermodynamic cycle to close, the contribution of these dummy atoms to the free energy of the system must be independent of the molecular environment. For example, if we delete a methyl group in a binding site and in solution, the free energy of the dummy methyl group must be the same in both environments. This criterion is met for methyl groups, because a dummy methyl group, in our simulations, is a simple set of masses and springs (and potentially torsions) with an energy that does not depend on the environment. Thus, any free energy contributions from the dummy atoms in the binding site rigorously cancel with contributions from the dummy atoms in solution, the other side of the thermodynamic cycle. Thus, the bonded terms of dummy atoms make no contribution to the relative free energy[10, 9].

In general, dummy atoms do not contribute to the relative free energy for all transformations of groups of atoms into dummy atoms except when the geometry is modified due to the influence of external forces [10, 9]. That is, cancellation will occur for any simple deletion of singly-connected groups of atoms. However, bonded terms *can* in some cases contribute to relative free energies under the influence of external forces[10, 9]. The main scenario in which this can be expected to happen is for deletion of multiply connected groups. For example, for mutation of cyclohexane into butane, dummy carbon atoms left behind from cyclohexane can affect the free energy in at least two ways. First, since the dummy atoms are still bonded to the butane atoms, they can affect the conformation of butane, altering which states are preferred. Second, the free energy of the system of dummy atoms (masses and springs) depends on the conformation of butane. If the preferred conformation of butane is different in complex versus in solvent (such as due to contacts with the receptor), the free energy of the dummy atoms will be different, producing a thermodynamic cycle which does not properly close due to free energy contributions from these bonded terms. This has the potential to happen whenever portions of rings (partial loops of atoms) are being deleted. Practical complexities such as the necessity to break or form bonds when opening or closing rings also make these cases difficult, as discussed elsewhere[10, 9].

So, we can avoid any contributions from the bonded terms of dummy atoms by avoiding breaking or forming rings, but in some cases this may be necessary, as we explain below. If ring breaking or forming is necessary, any contributions from dummy atoms will be small whenever the conformation of the molecule, with its dummy atoms, is essentially the same in water versus in complex, which will happen when the portion of the molecule being left behind is rigid. For example, for a naphthalene to benzene transformation or similar (Figure 4.1), benzene is quite rigid, and has a very limited range of motion regardless of environment. It is unlikely that a binding site could alter the conformation of benzene enough in order to induce a significant change in the bonded energies of the naphthalene dummy atoms left behind. In such cases, there may still be a small contribution of bonded terms to the overall relative free energy (that is, the thermodynamic cycle may not quite close formally) but we believe the effect will be quite small. In contrast, a naphthalene to cyclohexane transformation is much more risky, because the ring behind left behind now becomes flexible (cyclohexane) and its bond lengths will change appreciably, substantially affecting the bonded energies of the dummy atoms left behind from the additional ring system.

Here, then, our goal is to avoid inserting or deleting any partial rings as by so doing we can avoid introducing any errors due to bonded contributions to relative free energies. However, making this an absolute rule proves to be a bad idea. For example, if the set of molecules contains one group consisting of benzene derivatives (potentially with R-groups attached in various places) and another group based on the naphthalene scaffold, with two aromatic rings (again potentially with R groups) linked together. If we abide by the rule of never inserting or deleting partial rings, we will end up with two disconnected groups of molecules and no way to compare their binding affinities. We propose that in such cases, where rigid rings are being retained in a proposed partial ring deletion, it is acceptable to delete the partial ring in question and assume that bonded contributions to the free energy cancel (green arrow, Figure 4.1). So, our goal is to prefer relative free energy calculations which preserve rings or avoid deleting partial rings, but when absolutely necessary, we will tolerate deletion of partial rings as long as the components being left behind are essentially rigid.

Goal 5: Every molecule must be part of at least one closed thermodynamic cycle

Free energy calculations can yield accurate free energies, or results which are wildly wrong[14, 189, 37]. When the latter situation occurs, the source of error can be difficult to discern. Assuming the system being modeled is representative of experimental conditions, there are two main sources of error – poor convergence (i.e. the free energies would have been correct if only enough simulation were done) or force field inadequacies (i.e. additional simulation would have kept computed free energies the same and only reduced the uncertainty). One way to check for convergence problems is to add some redundancy into relative free energy calculations, introducing additional calculations between some molecules and hence adding some cycles – closed paths around which relative free energies must formally sum to zero. The difference from zero is called the cycle closure error. This has been done in some relative free energy calculations in the past (for example references [170, 30, 37, 173, 118, 133, 31, 181]) and provides a lower bound on the amount of convergence error in the calculations. The literature suggests that this can be a useful lower bound, in the sense that sometimes the cycle closure error is extremely large, as much as several kcal/mol [170, 30, 37, 173, 118].

One frequently requested feature from modelers in industry is the ability to know when calculations are expected to fail, and this provides at least some information in that regard.

Overall, then, in order to provide some level of consistency checking and detection of convergence errors, we require every molecule be part of at least one closed thermodynamic cycle.

Goal 6: The set of planned calculations should be spanned by relatively few calculations

When computing relative free energies across a large set of molecules, we may need to combine results of multiple calculations, leading to an accumulation of statistical error. To prevent these errors from becoming too large, we want to be able to get between any two molecules with no more than a certain maximum number of calculations.

3.3.2 Algorithm

Our main goal, then, is to construct a undirected graph where each node (or vertex) is a compound of interest and each edge (or arc) a relative free energy computation for the two flanking compounds. The edges should be assigned in a way to accomplish our design goals mentioned above. Effective free energy computation typically requires that the two molecules be sufficiently similar. So the edges (planned calculations) will depend heavily on the computed similarities between molecules. In the following, we introduce our definition of the *similarity* concept.

We want a similarity measurement to assess ease of computation

Intuitively speaking, similarity refers to the “likeness” of two molecules and is one of the factors that strongly influences the success of relative free energy calculations. For this reason, similarity is a pivotal concept in our algorithm, and we define it as a scalar quantity measuring the feasibility of a specific calculation. The higher the similarity, the more feasible the calculation is. To measure similarity, we seek scores spanning a known range, with the minimum representing total dissimilarity, and the maximum representing identity. We choose the range [0, 1] to denote the similarity score. This choice of the range is somewhat arbitrary but has several advantages described below.

This definition of similarity relates directly to how we construct the graph of planned calculations. When the similarity between two compounds is high, we should have a higher likelihood of connecting them by an edge. Often, the term “similarity” in the literature refers to a simple measurement of chemical similarity. But here we move past simple chemical similarity and define similarity scores which ensure our design goals will be met. Thus, we are more concerned with assessing the *computational* difficulty of particular transformations than rigorously scoring chemical similarity. That is, we are more interested in scoring computability. To appreciate this, consider a possible transformation between methane and ethane, and another transformation between benzene and toluene. Chemically, benzene and toluene are more similar, but in (in the absence of other sampling considerations) RBFE calculations between both pairs are expected to require roughly the same amount of computational resources because the mutation is identical for both pairs (hydrogen to methyl). Thus, to measure computability, similarity scores should be very close (if not identical) for methane-to-ethane versus benzene-to-toluene transformations.

To score efficiency of transformations based on a chemical similarity metric requires one major assumption – that the most important contribution to a transformation’s difficulty is the magnitude of the transformation itself. This will certainly not always be the case. For example, in a hypothetical receptor, one extremely small chemical modification (introduction of a methyl at a particular location, for example) might introduce a dramatic binding mode change or a new receptor conformation, posing substantial computational challenges, while a much more dramatic modification (introducing a phenyl group elsewhere in the molecule, for example) might do very little to the binding mode and receptor conformation and be computationally straightforward. However, in the limit of adequate binding mode and receptor sampling, transformation difficulty is an important metric. And even when facing potential problems in binding mode and receptor sampling, if we lack information on when to expect these effects, we should still be best served by focusing on the difficulty of particular transformations.

Our similarity measurement starts with the size of the maximum common substructure

A common theme of Goals 1-2 above (Section 3.3.1) is the desire to minimize the number of atomic insertions and deletions, which we can build in to our similarity scores by using the size of the maximum common substructure (MCS) as the foundation for scoring. An MCS search determines the largest common substructure shared by two molecules. Once this is known, the number of atomic deletions or insertions is immediately apparent, as is the change in any ring systems.

Here, we can make an MCS search better suit our needs by modifying it slightly. Specifically, we treat all heavy atoms as equivalent in the search, since as noted in Goal 1 (Section 3.3.1) , changes in atom types are straightforward. Thus our search focuses on molecular topologies rather than atom types. We also adjust the search to prefer substructure matches which preserve ring systems as much as possible (Goal 2). The resulting modified MCS search forms the foundation for our similarity scores.

Initial similarity scores are calculated based on the size of the MCS using the expression

$$S = \exp\left[-\beta \times (N_A + N_B - 2N_{MCS})\right] \tag{3.1}$$

where β is an arbitrary constant value, and N_A, N_B, and N_{MCS} are the number of heavy atoms of the two input molecules and of the MCS, respectively. Thus the term $N_A + N_B - 2N_{MCS}$ is the total number of atoms inserted or deleted in the transformation. We use an exponential both to ensure that scores range between 0 and 1, and to strongly favor small structural changes.

We construct similarity scores based on the modified MCS search, and fold in other scoring rules

We want similarity score calculations to be easy to automate and extend. As noted, the MCS similarity provides the foundation for our scores, but we still have several other goals (Section 3.3.1) to achieve. The easiest way to do this is by modifying our similarity score to include these aspects as well. Specifically, we need to also ensure that compounds being compared share the same net charge, and avoid broken or partial ring systems in our transformations. To include these factors in similarity scores, we built a simple *rule engine*, which provides a mechanism to combine multiple requirements (rules) or scores into a total. A rule here represents a regulation for similarity calculation, and its application involves taking in a pair of input molecules and outputting a similarity score based on the rule. The MCS search above can be recast in this format as a "maximum common substructure rule" (MCSR), which computes the size of the MCS and outputs a score in the range [0,1].

Our rule engine allows straightforward combination of multiple rules, typically by multiplication to maintain the score range. In addition to the MCSR, we also apply a "minimum number of common atoms rule" (MNCAR) which says that the two molecules must share at least n heavy atoms to be regarded as similar. This simple rule outputs a score of 1 if the number of common atoms is larger than n, and 0 otherwise. A composite of the MCSR and MNCAR rules amounts to checking the number of heavy atoms in the MCS, and if it is greater than n, returning a score based on the MCSR, otherwise returning 0. The score of the composite rule is given by the simple formula: $S = S_{MCSR} \times S_{MNCAR}$, and thus S remains a number in the range [0, 1] . This approach is useful because it decouples rule definitions from the scoring process, making it simple to add additional rules without modifying existing ones. For example, to add a new "equal charge rule" (ECR), which, given a molecule pair, outputs 1 if they have same the net charge and 0 if not, to we simply calculate S_{ECR} (the score for the equal charge rule) and add it to the composite rule, which is now using $S = S_{old} \times S_{ECR}$. For more details, interested readers are encouraged to read the LOMAP source code and the documentation therein.

Here, we are able to build Goals 1-4 (Section 3.3.1) into our similarity scores by forming a single composite score out of just four specific rules. These goals are handled as follows:

- Goal 1 (minimize atomic insertions and deletions): This is built into the definition of the MCS score itself (and the corresponding rule, MCSR), as described above.
- Goal 2 (preserve rings as much as possible): This, too, is built into the MCSR.
- Goal 3 (preserve net charge): The equal charge rule (ECR) zeros similarity scores for molecule pairs having different net charges.
- Goal 4 (avoid breaking rings, and only break them if planar): Here, we examine the MCS to see whether it contains broken or partial ring systems. If it does, we delete atoms in the broken or partial ring systems (and any unconnected moieties due to the deletion). Since this rule is based on the deletion from the MCS, we call it trimmed-MCSR (TMCSR) and calculate resulting score using the formula:

$S = \exp[-2 \times \beta N_{del}]$, where β is the same constant value as in the MCS rule, and N_{del} is the number of deleted heavy atoms. Here, we actually assign two different TMCSR scores, one we call "strict ring deletion" and a second we call "loose ring deletion". In both approaches, ring atoms in one molecule cannot be mapped to non-ring atoms in another molecule. The difference is in handling of joined ring systems, as in Figure 4.1. In strict ring deletion, if any component of a joined ring system no longer remains a ring after MCS calculation, the entire ring system is deleted. In contrast, in loose ring deletion, ring atoms in the MCS are kept if the portion of the ring system being left behind remains planar. The latter allows transformations like naphthalene to benzene, while the former does not.

Besides applying the above rules and constructing a composite rule by taking the product of their scores, we also define a simple cutoff rule (the minimum similarity score, "MSS"): Compare the similarity score with a threshold value, return 1 if it is greater than the threshold, otherwise return 0. The goal of this rule is to ensure that calculations in the final graph meet at least a certain minimum level of similarity. This rule is not used in construction of the composite similarity scores, and we discuss its application below.

Calculation planning builds up a graph based on similarity scores

We construct our graph of planned RBFE calculations following this procedure:

1. First, we calculate two sets of similarity scores from all pairs of the molecules. The first set, called the loose scores, are obtained by execution of a composite rule composed of the rules from Goals 1-3 and Goal 4 (loose). The second set, called the strict scores, are obtained by execution of a composite rules composed of the rules from Goals 1-3 and Goal 4 (strict). We store these two sets of scores in two separate matrices. The strict scores are used throughout except where otherwise noted.

2. We use the strict score matrix to assign an edge connecting any pair of nodes with a final similarity score greater than zero, and then we remove all the edges with scores less than the MSS cutoff, thus obtaining an initial graph. The following steps will refine this initial graph and reduce the number of edges.

3. The initial graph often has several connected components – subgraphs that have no connections between one other but within which there is a path between every pair of nodes. The more similar molecules are, the more likely they will be within the same connected component, and vice versa. We call all of the nodes within a connected component a cluster. Within each cluster, there are, at this point, typically far too many edges because most nodes are directly connected. We then try to reduce the number of edges while ensuring the cluster continues to meet our design goals (especially Goals 5-6, which are not encoded into the similarity scores), which we can think of as constraints. Specifically, we want the minimum number of edges such that 1) every node should be in a cycle, and 2) the length of the shortest path from any one node in the cluster to any other node in the cluster should be less than a specified maximum distance (*MAXDIST*). For 1), we use Paton's algorithm [135] to calculate the cycle basis [7], and use this to determine if all the nodes are in a cycle. For 2), we use a breadth first search to calculate the diameter of the cluster; from the diameter, we determine if the cluster meets the distance constraint. Following this procedure, if the constraints are met, we then sort and list the edges by their similarity score, from lowest to highest. We then remove the least similar edge and check if the cluster still meets the constraints. If it does, we then remove the next least similar edge, and so on, until we have checked every edge and there are no remaining edges that can be removed without violating the constraints. Thus, at the end of this process we obtain a graph with a minimum number of connections, which we call the minimized graph.

4. At this point, different structural clusters still remain disconnected. Therefore, we

use the loose score matrix to merge clusters from the minimized graph into a final, connected graph. This step has two passes. In the first pass, we connect as many of the clusters as possible, while in the second pass we build in cycles between clusters. The first pass connects clusters via a weighted maximum spanning tree. For the second pass, we repeat the procedure used in the first pass, but omit any edges created in pass one, thus creating a second weighted maximum spanning tree. Thus, each cluster will now be connected via two routes, ensuring our goal about cycles is met. It is important to note, however, that only connections with nonzero similarity scores are allowed, so clusters of different net charge will remain disconnected, as will any compounds or clusters which share no similarity with other compounds.

The MSS rule described above is applied when planning all intra-cluster calculations, but it is not applied at the stage of planning calculations spanning structural clusters, simply because we wish to ensure a path between all compounds of the same net charge if at all possible. Thus the final graph may have connections with similarity below the minimum similarity score, though this will only happen if it is necessary, since there is still a preference for connections with higher similarity scores.

Because *MAXDIST* (step 3) applies only within each structural cluster, the total number of calculations across the final graph is actually larger than *MAXDIST*, in a way that depends on the nature of the set of compounds. If there are m initial structural clusters, then the upper bound on the maximum distance across the graph after step 4 is $m \times MAXDIST + (m-1) = m(MAXDIST + 1) - 1$. In general we expect that for a typical congeneric series, the maximum distance across the final graph will end up being *MAXDIST* since all molecules will be in a single structural cluster. However, for more diverse compound libraries there may be several distinct structural clusters (such as for the "trypsin" dataset in our Supporting Information, which contains the Maybridge fragment library as a subset).

In some sets of planned RBFE calculations, it may be desirable to include a set of knowns – compounds having known binding free energies – both to allow calculation of absolute binding free energies of the unknowns by referring to the knowns, and as a consistency check. Our code allows the user to provide a set of known compounds which are a subset of the whole, and in this case, the maximum distance, *MAXDIST*, is set to apply to the distance between any given unknown and some known compound. So the calculation planning algorithm will ensure every compound is within *MAXDIST* of a known compound. This actually reduces the number of required calculations (since now unknowns are allowed to be more than *MAXDIST* apart).

3.3.3 Our implementation is in Python

We implemented the above algorithm in Python 2.7. The third party libraries used here include Schrödinger's [149] or OpenEye OEChem's APIs [132] for general molecular manipulation, MCS searches, and modification of MCS output (such as removing partial rings, etc.), networkx [60] for graph creation, traversal, and manipulation, OpenEye OEDepict's APIs [132] for molecule depiction and graphviz [38] for visualization of the graph.

Our toolkit is available through SimTk as Lead Optimization Mapper, LOMAP, at `https://simtk.org/home/leadoptmap`.

3.4 Results

In some sense, our key result here is LOMAP itself, which provides a general tool for automatically planning RBFE calculations. It also outputs the trimmed MCS for each planned calculation, making setup of the actual input files for RBFE calculations straightforward in at least several common simulation packages (DESMOND[148, 13] and GROMACS[68], for example). However, it is worth briefly assessing the output of LOMAP on some test sets.

Here, we applied LOMAP to several different sets of molecules, and full detail test sets and output are provided in the Supporting Information. Our main focus here is on a set of 50 Factor Xa (FXa) inhibitors taken from various literature sources[112, 186, 185, 150, 187], and the LOMAP plan for the full set is shown in Figure 3.2 (with structures for a smaller subset in Figure 3.3). But we also tested a set of potential trypsin inhibitors (a fragment library which was screened for binding to trypsin[129], as well as a substantial number of known trypsin inhibitors), and the SAMPL3 [126] set of small molecules. Results for these are provided in the Supporting Information, and statistics on the resulting graphs are shown in Table 1.

Here, we used LOMAP's default parameters, setting $\beta = 0.1$ and the similarity cutoff to 0.05. These parameters correspond to a cutoff of $(N_A + N_B - 2N_{MCS}) = 30.0$ in the MCS rule[1]. This means that we will not consider any calculations involving 30 or more heavy atom insertions or deletions. This specific choice is somewhat arbitrary and we plan on testing the specific choice using detailed free energy calculations in the future. At this point, our choice was made based on the assumption that for transformations larger than this, the expected error in computed relative binding free energies will be large and/or convergence will be extremely difficult. We also set $MAXDIST$ to 6, again somewhat arbitrarily, knowing that statistical error accumulates with each additional edge (and so larger $MAXDIST$ will introduce additional error) while at the same time smaller $MAXDIST$ requires substantially more computational effort. Again, systematic investigation of the optimal balance here will require a large number of free energy calculations, and is an important topic for future work.

In the FXa set, LOMAP automatically divides the molecules into 3 separate groups with different net charges. The total number of calculations is 65, only marginally larger than the number of molecules in the set (50), and much smaller than the 1225 possible pairwise

[1] assuming $N_{del} = 0$; if it is not, the threshold is adjusted slightly.

combinations of molecules. In general, the final product (Figure 3.2) meets our design goals, including building in closed cycles (Goal 5). (One specific cycle is highlighted *via* green lines in Figure 3.2). However, one node, 20523 (red in Figure 3.2) is not in a cycle. This ends up being because, due to the MSS rule, 20523 and 20577 are the only two members of one structural cluster. While 20577 is similar enough to other molecules that it ultimately gets connected to other nodes, 20523 is not, so it is left not belonging to a cycle. Similar cases are found in the trypsin dataset we examined. The final maximum distance across the graph is 9 (Table 1).

Table 3.1: Statistics of LOMAP plans for the sets examined

Dataset	number of nodes (n)	potential edges ($\frac{n(n-1)}{2}$)	planned edges	final maximum distance
FXa	50	1225	65	9
Trypsin	576	165600	785	23
SAMPL3	36	630	50	6

Table 3.2: Shown are properties of LOMAP plans for relative free energy calculations spanning the three different test sets used here. The number of nodes (molecules) is shown, along with the number of possible free energy calculations between these molecules. Planned edges is the number of planned calculations spanning the set, and the maximum final distance is the maximum distance between any pair of molecules across the final graph.

Because the full set of molecules is still relatively large, even for FXa, it is worth examining a more detailed version of the map, along with chemical structures, to see whether the final graph makes intuitive sense. Figure 3.3 focuses on the portion of the graph with compounds having a net charge of +1. LOMAP automatically generates similar graphs, which contain some additional information (such as the trimmed common substructures), though these are too large to show here and samples are shown in the Supporting Information.

Some features of the FXa subset shown in Figure 3.3 highlight advantages of LOMAP over planning calculations manually by inspection. For example, the blue node in the center is selected by LOMAP to essentially serve as a hub (particularly highly connected), because it is the structure common to the largest number of molecules in the series. Having this scaffold as a hub dramatically reduces the distance between other compounds sharing the same scaffold. Furthermore, a manual search might propose a calculation between the compounds shown in orange circles. However, these, though sharing substantial similarity, have bicyclic rings of different sizes, which (because we seek to avoid breaking rings) would involve larger mutations – a transformation would involve deleting and reinserting both bicyclic rings. Instead, our algorithm connects these compounds by passing through the purple node, which involves modifying only one bicyclic ring at a time. Observations of this type highlight how automatic planning algorithms are essential for large scale relative free energy calculations, as planning at this level of detail would be impossible on large sets of molecules since there are simply too many possibilities to consider (see the trypsin dataset in the Supporting Information, for example).

Another important result of these calculations is the common substructure for each planned calculation, and the mapping of ligand atoms onto this common substructure, which is also an output. This information makes it simple to set up what we call "single topology, explicit intermediate" binding free energy calculations (Figure 3.4). In these calculations, each ligand is perturbed to the common substructure, both in the binding site and in solution. From the difference in free energies one can obtain the relative binding free energy. These calculations are conceptually (and algorithmically) typically very straightforward to set up, since they involve simply turning any atoms being "deleted" into dummy atoms and making any changes to atom type (and corresponding bond, angle, and torsion parameters) needed to get to the common substructure. Unlike calculations going directly between ligands, these do not require a mapping of atoms from one ligand onto the other, nor do they require simultaneous transformations to and from dummy atoms, which simplifies setup. We are working on automated tools to set up such calculations in some common simulation packages, and plan to release these separately.

3.5 Discussion and conclusions

LOMAP provides an automatic way to plan relative binding free energy calculations, which has been one major hurdle hampering more widespread application of these calculations to problems in drug discovery. As input, it takes a set of potential ligands of interest, and outputs a map of planned free energy calculations spanning the set with relatively few transformations which are designed to be relatively efficient, based on the number of atomic insertions and deletions required. This map also has several other features designed to aid overall accuracy and provide consistency checking. Specifically, it keeps overall distance across structural clusters below a specified threshold, and it builds in closed cycles of mutations to allow consistency checking and provide additional information about when the calculations may be performing poorly. Our approach also folds in several other major considerations for accuracy, such as avoiding calculations between molecules of different net charge, proper handling of partial ring deletions, and so on.

LOMAP provides a first effort at systematically planning efficient free energy calculations. Much follow up work needs to be done to determine whether the choices make here actually lead to the most efficient free energy calculations, or whether other choices are better. While here, we have focused on minimizing the number of atomic deletions and insertions (which is supported by the free energy literature) one might imagine other criteria might also be important measures of the efficiency of potential relative free energy calculations. For example, swapping one bioisostere for another might be preferable even if it results in a larger number of deletions or insertions. Or a transformation which preserves the topological location of hydrogen bond donors and acceptors might be preferable over one that does not, even if it involves a few more deletions and insertions. So far, we are not aware of any efficiency data from free energy calculations which sheds light on these issues, so the current implementation is like a good starting point. But further study is needed. Hopefully the approach presented here will provide the foundation for new systematic studies of transformation efficiencies.

Since we hope this will be the foundation for much further development in the area, our code is open source, under the BSD license. We have designed the graph planning algorithm itself to be modular, taking a set of arbitrary similarity scores as input, so that the planning component can be easily modified and extended. Also, our rule engine is designed to allow easy incorporation of additional rules, and/or replacement of existing rules with new ones.

Overall, we believe this approach provides a promising way to begin automating the setup of large scale relative binding free energy calculations. As noted above, with the output of these calculations – the plan of calculations and the common substructure for each planned calculation – it is simple to automate setup of input files for many common simulation packages which perturb each ligand in a pair to the common substructure. Thus we hope that Lead Optimization Mapper will pave the way to applying binding free energy calculations on a larger scale in a wide range of applications.

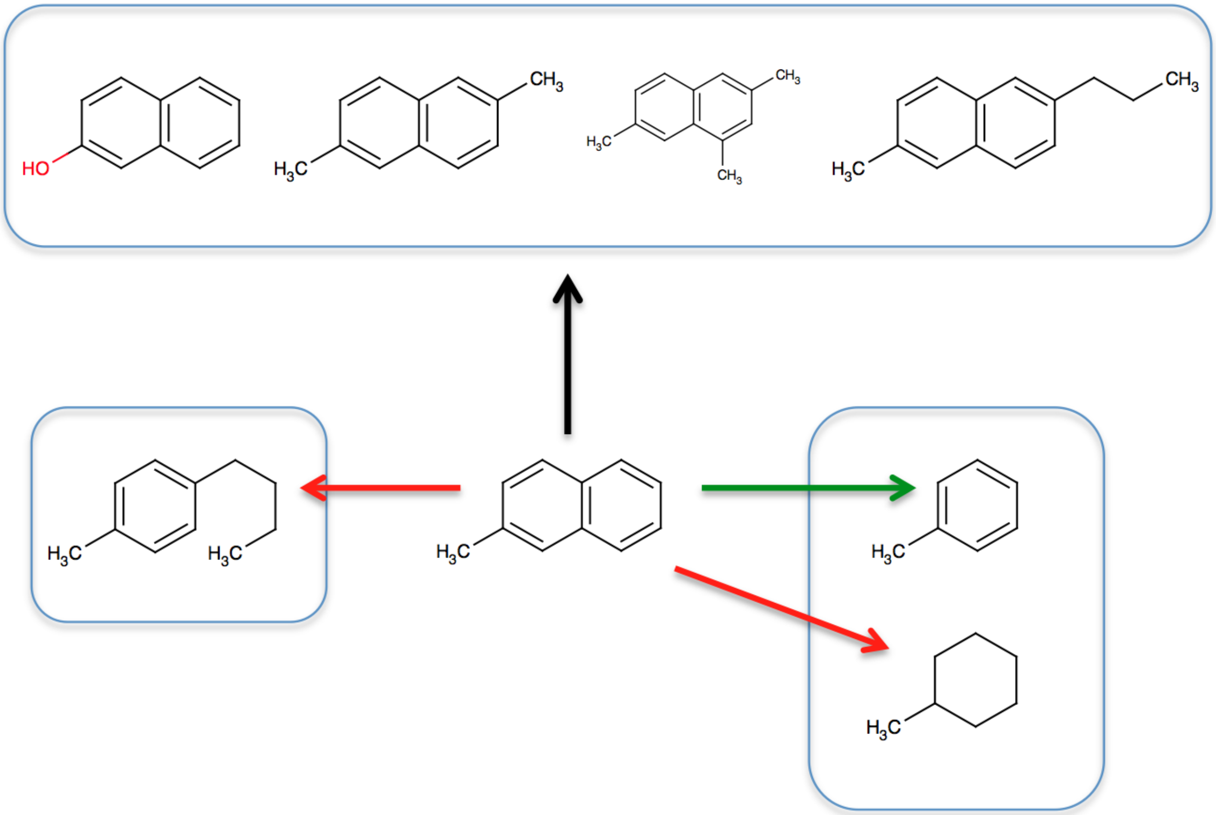


Figure 3.1: Favorable mutations of 2-methylnaphthalene, based on our design goals. Black and green arrows show allowed transformations from 2-methylnaphthalene. Red arrows show transformations which are not allowed. Transformations marked by the black arrow are considered the most favorable, and similarity (and hence favorability) decreases from left to right, based on design Goal 1. Mutation from 2-methylnaphthalene to 1-butyl-4-methylbenzene is prohibited (red arrow to left) because it would involve breaking a ring in a bi-cycle (Goal 4). Mutation to toluene (green arrow) is allowed only under the "loose" scoring scheme, and only if necessary to span the set (Section 3.3.2), while mutation to methylcyclohexane is prohibited (red arrow to right) because it involves breaking a ring in a bi-cycle and leaving a flexible ring behind (design Goal 4).

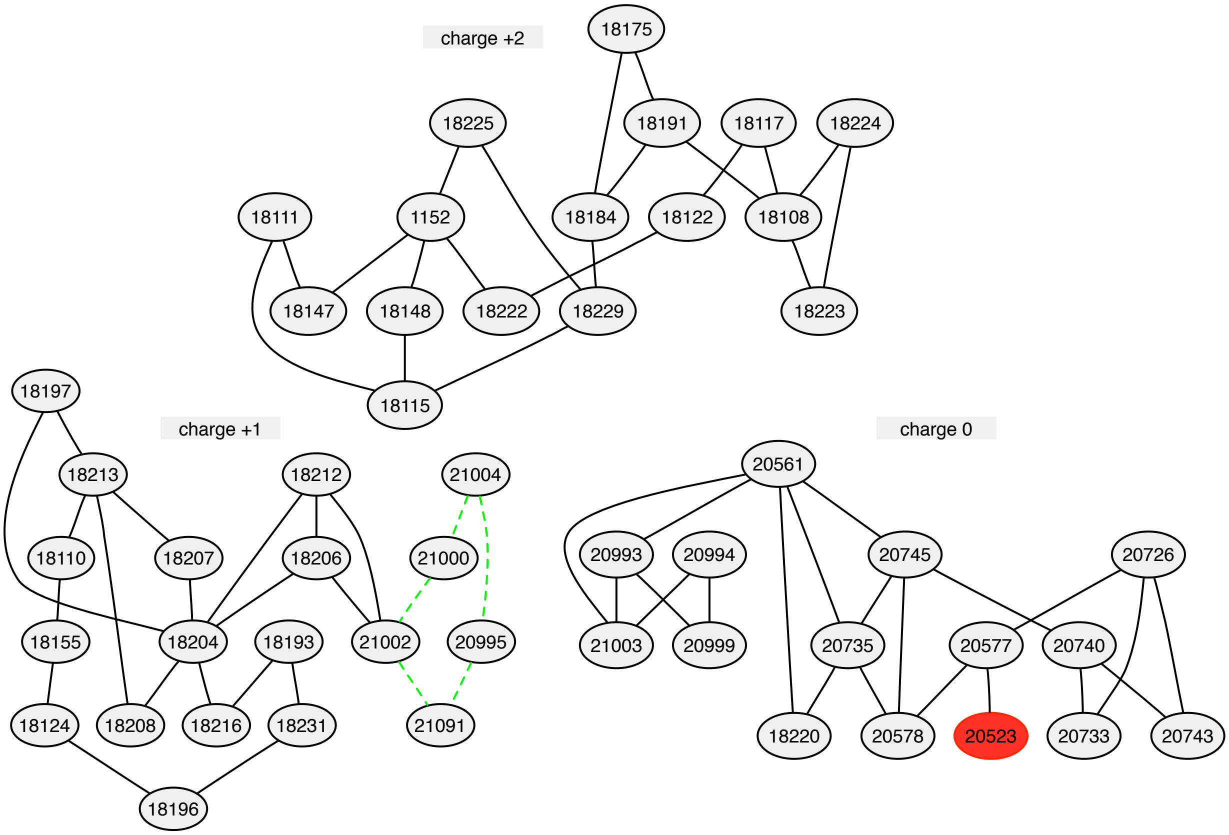


Figure 3.2: Planned RBFE calculations for the FXa set. Ovals represent molecules, the number inside nodes is molecule title, lines represent planned RBFE calculations. Green lines are a cycle example. Every nodes are in cycle except the red one.

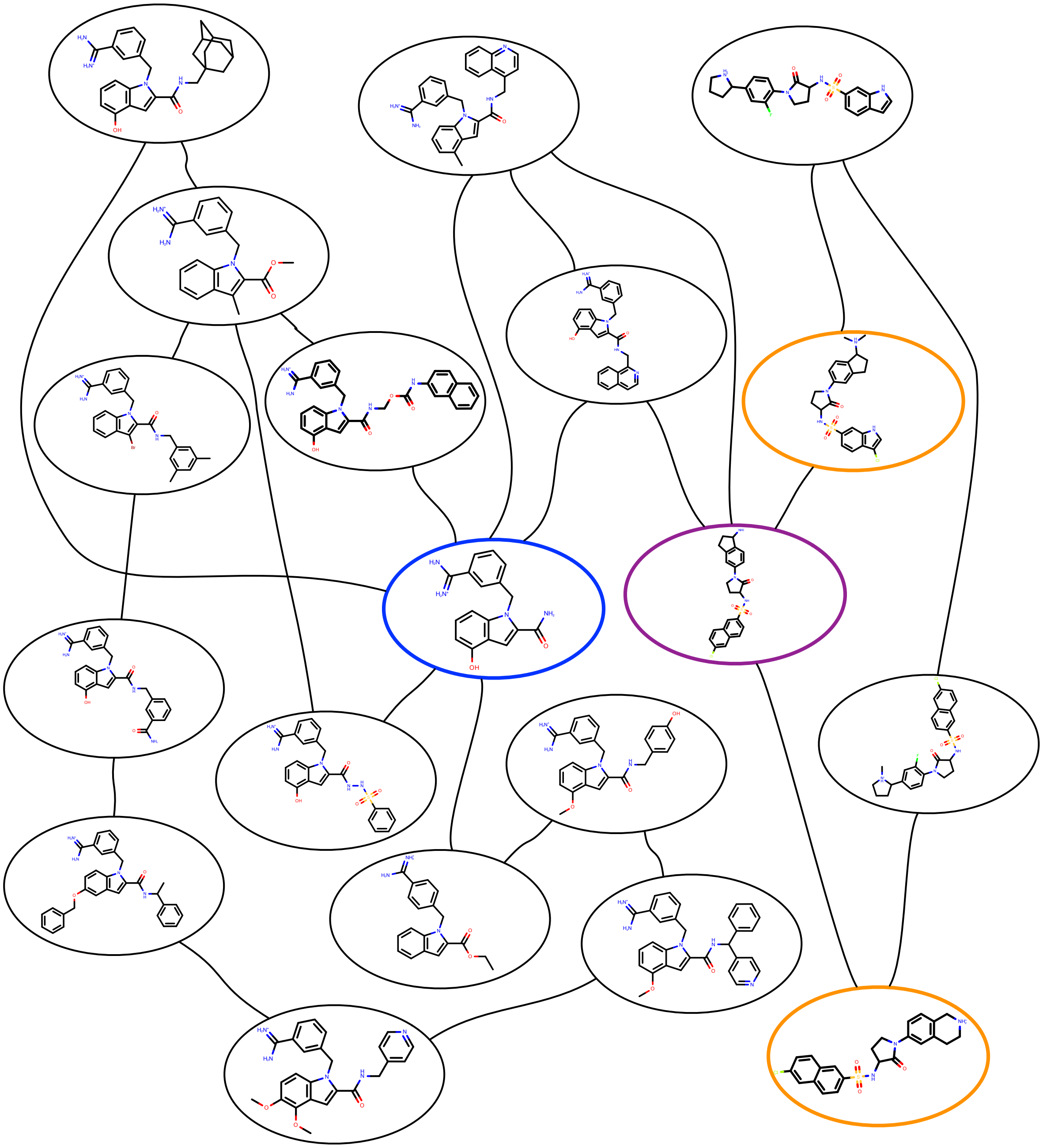


Figure 3.3: Planned calculations for the charge +1 subset of the FXa test set, showing details of molecular structures. Molecules highlighted in colors illustrate features of the output discussed in the main text.

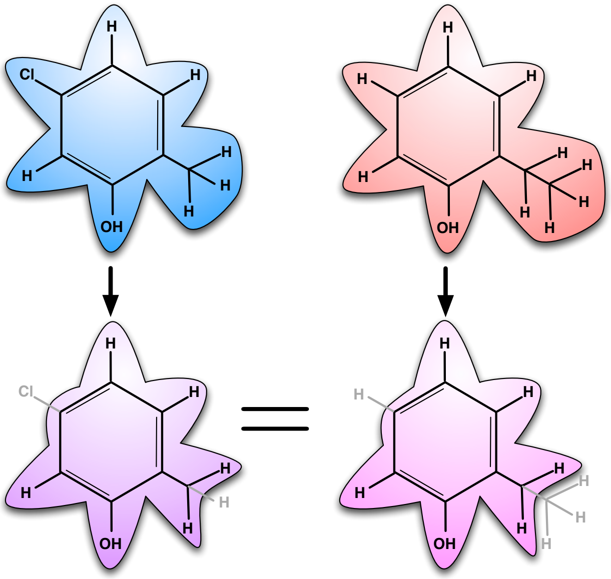


Figure 3.4: Single topology, explicit intermediate free energy calculations. Here, these calculations would be used to compare binding of 5-chloro-2-methylphenol and 2-ethylphenol. Fully interacting atoms are shown in black with underlying shaded contours, while noninteracting atoms (dummy atoms) are shown in gray with no shaded contours. The intermediate (scaffold) is specified explicitly, at the bottom. At left, the chlorine atom and one hydrogen are changed into dummy atoms, while at right, one hydrogen atom and a methyl group are changed into dummy atoms. The two scaffolds at bottom differ in number of dummy atoms, though these contributions cancel when computing free energies. The free energy calculation involves turning the specified atoms into dummy atoms in both molecules.

Chapter 4

Is Ring Breaking Feasible in Relative Binding Free Energy Calculations?

4.1 Abstract

Our interest is relative binding free energy (RBFE) calculations based on molecular simulations. These are promising tools for lead optimization in drug discovery, computing changes in binding free energy due to modifications of a lead compound. However, in the "alchemical" framework for RBFE calculations, some types of mutations have the potential to introduce error into computed binding free energies. Here, we explore the magnitude of this error in several different model binding calculations. We find that some of the calculations which involving ring breaking have significant errors, and this error is especially large in bridged ring systems. Since the error is a function of ligand strain, which is unpredictable in advance, we believe ring breaking should be avoided when possible.

4.2 Introduction

Here, our interest is in drug lead optimization, where a compound is known which binds the desired target, and we seek to create derivatives of this lead compound which either improve affinity or maintain affinity while improving other properties. Relative binding free energy calculations (RBFE) based on molecular dynamics (MD) simulations can be used to predict binding free energy differences based on chemical changes, in advance of synthesis of the derivative compounds. Thus, they can potentially substantially accelerate the lead optimization process [125] and are of considerable interest for drug discovery applications.

Given a particular model (force field and parameters), alchemical RBFE calculations yield correct relative binding free energies in principle, at least in the limit of adequate sampling, as reviewed elsewhere [117, 126, 24]. However, large chemical modifications require substantially more sampling, and hence, with a fixed amount of sampling, increasing the size of the transformation can increase the magnitude of errors due to sampling. Thus, in order to ensure typical modifications are relatively small, we recently designed a program, lead optimization mapper (LOMAP) [102], for planning efficient RBFE calculations.

LOMAP automatically selects single-topology RBFE calculations spanning a lead series by pairing similar molecules. In LOMAP, we only calculate the RBFE between molecules which have sufficient similarity. Structural similarities are computed on the basis of a similarity score, which relates to the change in the number of atoms during the transformation between the two molecules in question. Specifically, we identify the maximum common substructure shared by two molecules, and identify the change in the number of atoms needed to reach this substructure; we use these changes as the basis for our similarity score. Currently, LOMAP uses two scoring schemes, called "strict" and "loose", which differ only in how we treat transformations which would break polycyclic ring systems (Fig. 4.1).

In the strict scoring scheme, we do not allow any ring breaking when we search for the

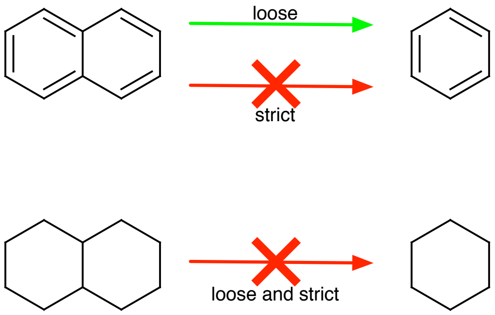


Figure 4.1: In the strict approach, mutation involving breaking any ring is not allowed (red arrows). Here, “ring breaking” refers to any transformation where atoms remaining in the final system were part of any ring which has been removed. In the loose approach, ring breaking mutations are allowed only when the ring system left behind is rigid (green arrows).

maximum common substructure. For example, in the strict approach, if we consider the pair naphthalene and benzene (Fig. 4.1), these have no common substructure because a naphthalene to benzene transformation would involve breaking a ring. On the other hand, in the loose scoring scheme, we allow ring breaking happen only when the ring system left behind is relatively rigid or planar (typically aromatic). For example in this case, mutation from naphthalene to benzene is allowed because the remaining ring, benzene, is rigid, while the mutation from decalin to cyclohexane (Fig. 4.1) is not allowed, because the remaining ring, cyclohexane, is not rigid (indeed, it can undergo significant conformational transitions). LOMAP was designed to use the loose scoring scheme only when absolutely necessary in order to produce RBFE calculations spanning a lead series – for example, if a group of bicyclic molecules could not be connected to single-ring systems via any other means – but it avoids these types of transformations whenever possible. This was done because the effective “deletion” of partial rings in polycyclic (bicyclic, in this example) ring systems can introduce error into the associated thermodynamic cycles(Fig. 4.2).

Effectively, the loose scoring scheme means that the thermodynamic cycle our RBFE calculations are based on in these cases no longer formally closes; we potentially have a missing contribution due to a conformational change in the remaining dummy atoms induced by changes in the conformation of the connected interacting atoms, shown by the red “approx-

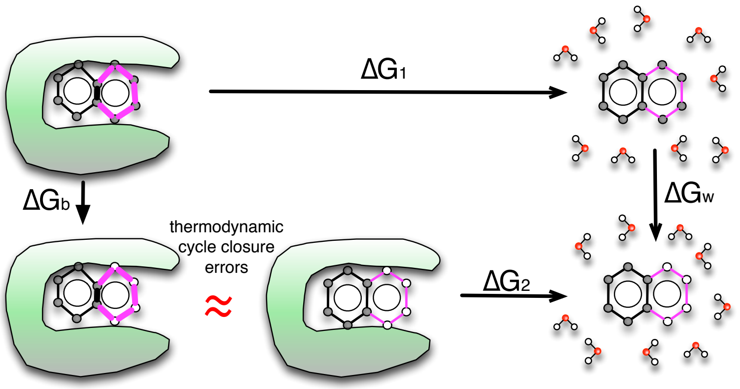


Figure 4.2: Here, we consider the thermodynamic cycle used for single-topology relative binding free energy calculations, where we compare binding of two ligands 1 and 2 (top and bottom, respectively, where our example shows cartoons of naphthalene and benzene). In this approach, we obtain the difference between the binding free energy of 1, ΔG_1, and the binding free energy of 2, ΔG_2, by perturbing the first ligand into the second in the binding site (ΔG_b) and in water (ΔG_w). These perturbations can involve changing some of the atoms in one or both ligands into *dummy atoms* (hollow spheres) – atoms which no longer have charge or Lennard-Jones interactions with the remainder of the system, but retain their bonded interactions. In a correct thermodynamic cycle $\Delta G_1 + \Delta G_w - \Delta G_2 - \Delta G_b = 0$. However, this is only strictly true when the free energy of the dummy atoms is identical in the binding site and in solution. This criterion is met when the dummy atoms are a set of masses and springs with a conformation which is not affected by that of the remaining "real" atoms. However, if the conformation of the dummy atoms is coupled to the conformation of the remaining portion of the molecule (i.e. when they are members of a ring), these two free energies may in fact differ and the thermodynamic cycle will not close. Here, at bottom left, the binding site is shown to introduce strain into the dummy atom ring (bold bonds), so that the free energy of the left bottom and middle bottom states are no longer equal, causing a cycle closure error (red "approximately equal" sign).

imately equal" sign in Fig. 4.2.

More rigorously, we show in the appendix that the dummy atoms can have a nonzero effect on the free energy change whenever there is more than one bond-stretch interaction between the same group of dummy atoms and the remaining interacting atoms (i.e. more than one connection point between the dummy atoms and the rest of the molecule).

In our previous work [102], we assumed that any contribution to the free energy change from these dummy atoms would be small in the case where the remaining atoms are in a rigid ring system, and larger when these atoms are in a flexible ring system. Thus, we assumed that, when necessary in order to ensure all compounds in a lead series could be connected, we could break rings in rigid molecules and still introduce only a minimum amount of error in computed binding free energies. However, this was an assumption – the exact magnitude of these contributions is not known. The existence of these errors was well known, but understanding their magnitude is now essential. This has dramatic implications for how we plan free energy calculations. Specifically, can we to allow these types of transformations in special cases? Or do they need to be avoided, or implemented via another route such as absolute free energy calculations? Here we aim to answer these questions. In our initial implementation of LOMAP we assumed that mutations beyond the loose scoring scheme, such as mutation from decalin to cyclohexane, will accumulate substantial errors and in general be unreliable.

Here, we compute the error introduced into single-topology RBFE calculations via both the loose and strict scoring schemes in several model binding calculations, as a function of the amount of strain in the remaining atoms in the ligand. These errors are limited to single-topology approach of the RBFE calculations, other methods like dual-topology approach [11, 136, 119, 115] and separated-topology approach [146] should not suffer from these errors because they handle these transformations without introducing multiply-connected dummy atoms which interact with remaining atoms. But since the single topology approach is the default method for LOMAP and is widely used in RBFE calculations, these errors are still important.

4.3 Method

As discussed in the Introduction, we test RBFE calculations in two bicyclic systems: (1) the transformation of naphthalene to benzene (Fig. 4.3); and (2) the transformation of decalin to cyclohexane (Fig. 4.4). We also test another case involving a bridged or cage-like ring system, system (3), the transformation of adamantane to bicyclo[3.3.1]nonane (Fig. 4.5). Here, to avoid the complexity and potential sampling problems introduced by doing these calculations in a receptor binding site, we model “binding” as the transfer of a ligand from water to a “binding site” consisting of the ligand in gas phase with conformational restraints which introduce strain.

This is sufficient for our purposes here, as we are solely interested in how introduced strain affects the formal accuracy of RBFE calculations —that is, we seek to determine how much apparent cycle closure error is introduced by ligand strain. Thus, our approach is sufficient, since the error is a function only of the difference in free energy of the dummy atoms when the ligand is under strain (Fig. 4.6). Since the dummy atoms do not interact with the rest of the system, the error is independent of the environment and dependent only on the degree of ligand distortion or strain.

In order to know how much error is introduced by the approximation in question, we need a way to determine the correct “binding” free energy of the molecules in question. We achieve this by computing the absolute binding free energies for every compound considered (Fig. 4.6). Here, we compute the free energy to move each molecule from water to the “binding site” —the absolute binding free energy —and then subtract these to obtain the relative binding free energy $\Delta\Delta G_{ab}$. Since absolute binding free energy calculations do not involve any ring breaking, they are correct, and provide the gold standard to which we can compare our relative binding free energy results.

Secondly, we calculate the relative binding free energy using a typical thermodynamic circle (Fig. 4.6) where we calculate the free energy change by transforming molecule 1 to molecule 2 both in water and in the binding site. From this, we obtain the relative binding free energy $\Delta\Delta G_{rl} = \Delta G_w - \Delta G_b$. Since this process includes ring breaking, $\Delta\Delta G_{rl}$ is our target result.

After doing both of these calculations, we compute the overall error as the difference between the reference result (from absolute calculations) and the target result. This measures the difference between the correct free energy change (as determined by absolute calculations) and the free energy change calculated by the relative calculations, which are in error (due to neglecting the free energy associated with a conformational change in dummy atoms) to some degree.

Here, we are interested in determining how the error changes as a function of ligand strain. Particularly, in the limit of no conformational change within the ligand, the conformation and free energy of the dummy atoms will be identical in both the binding site and solution, and no error will be introduced by deleting part of the ring. However, in the limit of very large strain, the bonds between the dummy atoms will be strained in the binding site but not in solution, introducing substantial error. To examine these effects, we added an additional bond between two atoms (bond type 6 in GROMACS), a spring, which is shown in bold in Fig. 4.6. By changing the length and force constant for this additional bond, we can then control the bond length for the shared bond for bicyclic systems or the distance between two end atoms in the cage-like system, which controls the strain in bonds between the remaining dummy atoms. The bond length details are shown in Tables 1, 2 and 3.

For each bond length we calculate the cycle closure error as described above. We find that substantial errors are introduced when the bond length change (strain) becomes sufficiently large. To provide perspective in terms of how much strain typically is introduced upon ligand binding, we examined simulations of several different protein-ligand systems and measured how much bond length change is typical on ligand binding.

For all the simulations, we used GROMACS 4.6. [68]

The initial structure files were generated by MarvinSketch 5.11.3 and then converted to mol2 files using the OpenEye OEChem toolkits [132]. The OpenEye OEChem Python toolkit and

Omega [66] were used to generate 3D conformations and assign AM1-BCC [78, 79] partial charges. Antechamber [178] from AmberTools 13 was used to assign GAFF atom [179] types and then AmberTools' tleap was used to generate the Amber prmtop and crd files which were converted to GROMACS format using acpype [160]. Small molecules were then set up in GROMACS and, for the solute-in-water case, solvated in TIP3P [82] water in a dodecahedral simulation box with at least 1.2 nm from the solute to the nearest box edge. The number of water molecules was 690 for benzene, 554 for napthalene, 678 for cyclohexane, 552 for decalin, 553 for adamantane and 597 for bicyclo[3.3.1]nonane.

AMBER combination rules (arithmetic average for σ and geometric for ε) were used. Simulations were run using Langevin dynamics, as previously [93], and the simulation timestep was 1 fs. Lennard-Jones interactions were gradually switched off between 0.9 and 1.0 nm, and an analytical correction was applied to the energy and pressure. PME was used for electrostatics, as previously, with a real-space cutoff of 1.2 nm. LINCS was used to constrain bonds to hydrogen. Each system and λ value (where λ is a parameter ranging between 0 and 1, where 0 corresponds to the unmodified system, and 1 corresponds to the end-state of the transformation) was independently minimized for up to 2500 steps of steepest-descents minimization.

Following constant pressure equilibration, box sizes were adjusted at each λ value by an affine transformation to ensure each λ value had the correct volume for the target pressure. After this, we conducted an additional 5 ns of constant pressure production simulation at each λ, discarding the first 100 ps as additional "equilibration", as previously [93]. Here we use Parrinello-Rahman barostat to modulate the pressure.

The parameter λ controls the transformation between end states. In this version of GROMACS, we use three separate λ values, one controlling modification of partial charges (λ_{chg}, turning solute partial charges to zero), the second controlling modification of the bond inducing strain (λ_{bd}, introducing this bond) and the third controlling modification of Lennard-

Jones interactions (λ_{LJ}, turning solute LJ interactions to zero). The details of the λ spacing can be found in SI.

In the case of the bridged ring system, we have to deal with an additional complexity. Because of the absence of a bridging atom in bicyclo[3.3.1]nonane, the internal non-bonded interactions involving atom A and atom B shown in Fig. 4.5 are different in adamantane compared to those in bicyclo[3.3.1]nonane - that is, the interactions differ not just in strength but in terms of which atoms interact. This is because the bridging atom changes which interactions are excluded and which are 1-4 interactions. Thus, the end state of the simulation which starts from adamantane (ΔG_w (Fig.4.7)) has the different non-bonded interactions from the starting state of the simulation beginning with bicyclo[3.3.1]nonane (ΔG_2 (Fig. 4.7)). Unless accounted for, these differences in non-bonded interactions will make the thermodynamic cycle fail to close even the absence of strain/conformational change, since we neglect a contribution due to the free energy of changing the internal non-bonded interactions. We call errors introduced by this change in internal non-bonded interactions "non-bonded discrepancy errors (NDE)". The NDE is not what we are interested in here, and also is not a necessary feature of binding free energy calculations - particularly, if our simulation package allowed us to change the exclusions and pairs lists with λ so as to remove the effects of the presence (or absence) of the bridging atom on 1-4 and excluded interactions, then we could compute relative binding free energies which were unaffected by NDE. Thus, we are interested in understanding errors *aside* from NDE. So, to avoid NDE errors in GROMACS, we modify the pairs and exclusions sections in our topology files to create an new reference molecule which has the same internal exclusion and 1-4 interactions as adamantane but the same atoms as bicyclo[3.3.1]nonane. This allows us to maintain the same exclusion and 1-4 interactions while transforming between a molecule which is like bicyclo[3.3.1]nonane into adamantane. With these adjustments, the simulation ΔG_2 has the same 1-4 interactions and exclusions as the simulation of ΔG_w; we call this case "adamantane-bicyclononane". As a comparison, we still run simulations without any adjustments to the topology file

(ΔG_{2_3b}). This case - which *does* include NDE - is called "adamantane-bicyclononane with NDE". adamantane-bicyclononane with NDE is analyzed using the same simulations as adamantane-bicyclononane (we use the same trajectory file and modify the topology file in order to evaluate the desired free energy using different interactions) as discussed in Fig. 4.7 (left bottom green/dashed green arrow in Fig. 4.7).

For all our systems, we use three different sets of special bonds to induce strain in the ligand. These vary in their bond length and force constant, and involve: (1) varying only the length (keeping the same force constant as the original bond, or as a normal carbon-carbon single bond in the case of the cage-like system (which does not initially have a bond between the shared atoms)); (2) varying only the force constant (keeping the same distance between the shared atoms as the original distance; in this case, the initial force constant is that of the original bond or a normal carbon-carbon single bond); (3) varying only the force constant but starting with a reduced distance of 88.7% of the original distance between the shared atoms (this is otherwise the same as case 2). With these combinations, we vary the distance between the shared atoms in the simulations over a wide range. We have 8, 5, and 5 simulations for sets 1, 2 and 3, respectively for the planar ring systems. In the adamantane-bicyclononane case (except for the NDE case) we add additional 5 simulations for set 2 to get a better coverage of the space of the bond length.

4.4 Results

Here, we examine how ligand strain in our model "binding" system impacts error in computed relative free energy calculations. Strain is controlled by an artificial bond which changes a bond length or distance within the ligand as it binds. In our bicyclic systems, we find that as this shared bond deviates from its original value (i.e., the ligand becomes more strained on binding) the error in the computed binding free energy increases. In the bicyclic systems (Fig.

4.3, Fig. 4.4) for the region close to the normal (unrestrained) bond length, the errors for both systems are relatively small and essentially statistically indistinguishable from zero – smaller than 0.5 kT.

On the other hand, for the cage-like system (the bridged ring case), errors are bigger. Unlike the bicyclic ring systems, here there is no bond between atom A and atom B in bicyclo[3.3.1]nonane (Fig. 4.5), making the distance between atom A and atom B differ substantially from that in adamantane. That is, the distances between atom A and atom B in the simulations of the absolute free energy calculation of bicyclo[3.3.1]nonane (ΔG_2 in Fig. 4.7) and the relative free energy calculation in vacuum (ΔG_b in Fig. 4.7) are significantly different.

This is not necessarily a problem - it just means that if we want to examine the error as a function of the distance between atom A and atom B, we have two different distances we can use, one which is substantially longer than the other. Thus, we plot the errors vs the distance between atom A and atom B in the simulations based on both of the references – 1, absolute simulations starting from bicyclo[3.3.1]nonane and 2, relative vacuum simulations starting from adamantane. We find that if we use the bond length in the relative calculations from adamantane in vacuum as the "original" bond length, the error is $\sim$ 30 kcal/mol when the bond length is 99% of its original value (Figures showing this result are in SI), while if we use the bond length seen in the absolute calculations from bicyclo[3.3.1]nonane as the "original" bond length, the error is $\sim$ 1 kcal/mol (4.5). For both of these simulations when the bond length changes by 1%, the errors are significant – larger than 1 kT. For adamantane-bicyclononane with NDE, with numbers which include NDE, compared with adamantane-bicyclononane, the errors are similar when the changes of the bond length are small and larger when the changes of the bond length are large (SI). This was expected because adamantane-bicyclononane with NDE includes additional errors beyond those in adamantane-bicyclononane – in addition to including contributions due to changes in strain/bonded energies, it also includes NDE errors (Fig. 4.7).

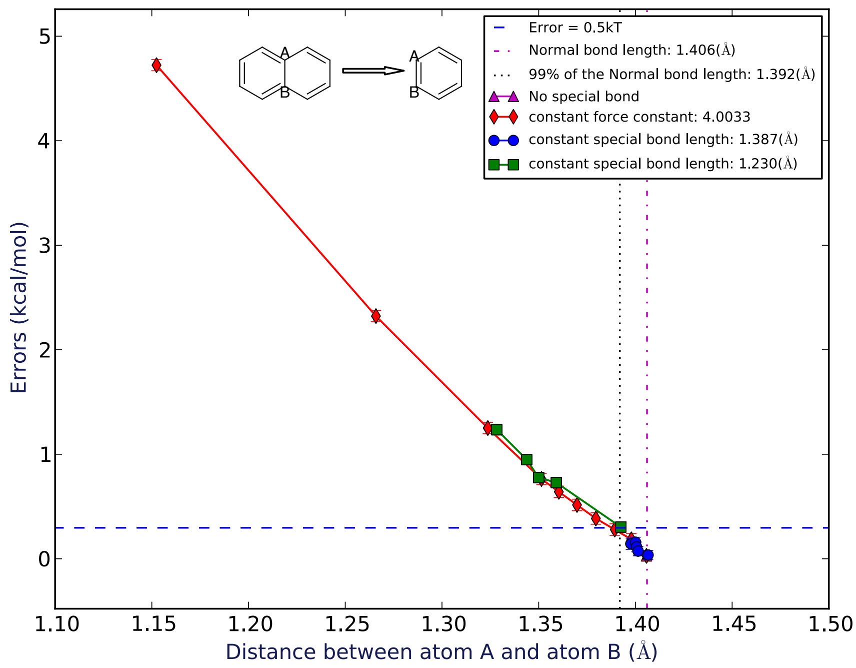


Figure 4.3: Errors for mutating naphthalene to benzene as a function of the average distance between the shared atoms observed in the simulations. Spots in different colors represent different force constants/bond length groups of the special bond.

Thus, we find that for the bicyclic systems, large bond length changes do lead to significant errors while small bond length changes (less than 2%) do not result in significant errors in relative binding free energy calculations. But for cage-like systems, even very small changes in internal distance (1%) can lead to very substantial errors (Fig. 4.5). Here, we assess significance based on the point at which the absolute error in the computed relative binding free energy becomes larger than the statistical uncertainty in our calculations.

For the bicyclic systems, we also examined the amount of strain induced by these bond perturbations in order to provide scale. We calculated the average energy difference between the most strained conformation at which the error is still statistically indistinguishable from zero (the “maximum indistinguishable” or “MI” case) and the original, unstrained case (“original”), both for relative calculations in vacuum. For the naphthalene to benzene system

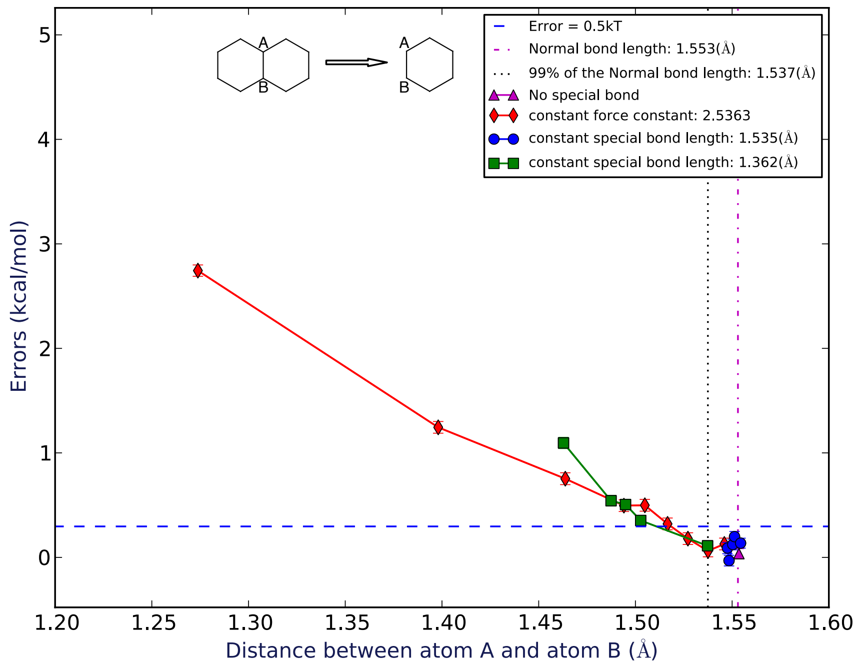


Figure 4.4: Errors for mutating decalin to cyclohexane as a function of the average distance between the shared atoms observed in the simulations. Spots in different colors represent different force constants/bond length groups of the special bond.

(Table 1), case MI is labeled with ID 1. The average potential energy difference between these two cases is 0.31 kcal/mol. For the decalin to cyclohexane system (Table 2), case MI is labeled with ID 2. The average potential energy difference between these two cases is 1.10 kcal/mol.

As noted, we originally expected that for the bicyclic system the flexibility or rigidity of the remaining ring system would have a substantial impact on the magnitude of the error, with rigid rings having substantially smaller errors than flexible rings. However, this is not what we find here —both approaches seem to have roughly comparable errors. However, we do find that for the flexible cage-like molecule, errors on ring breaking are much more substantial.

One possible explanation of this phenomenon is that, in the bicyclic system, the remaining

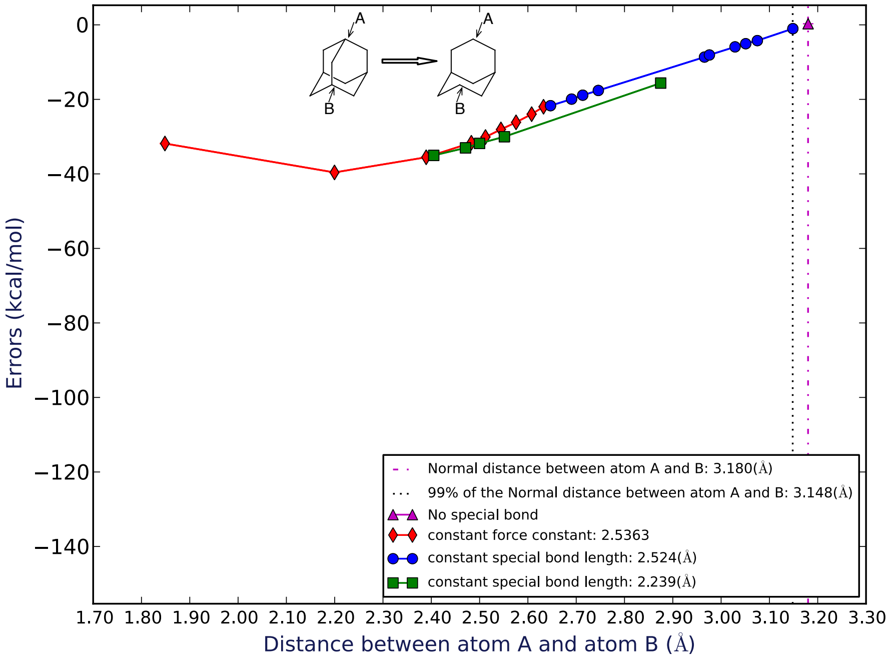
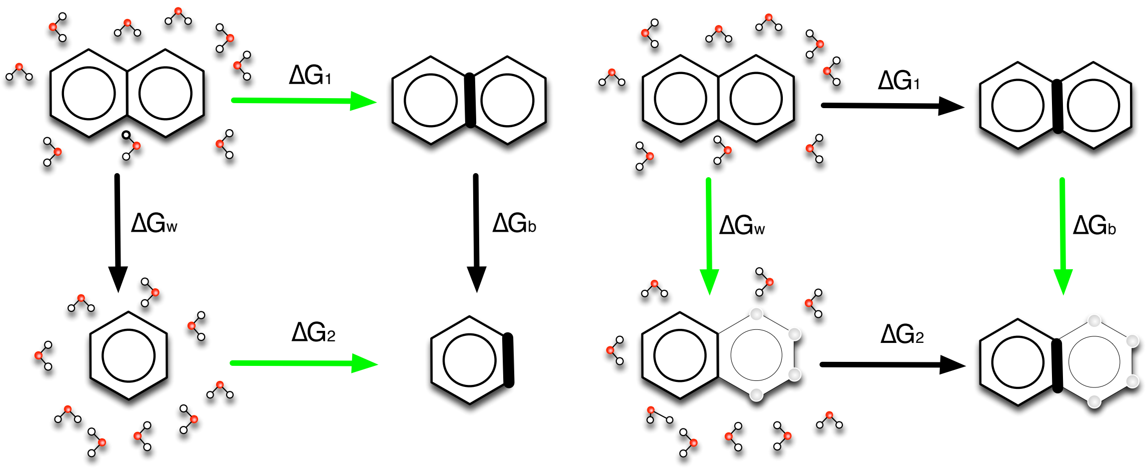


Figure 4.5: n
onane] Errors for mutating adamantane to bicyclo[3.3.1]nonane as a function of the average distance between the shared atoms observed in the simulations. This error is for the system of adamantane-bicyclononane without NDE. The ”original” distance is the distance between atom A and atom B in the absolute free energy calculation of bicyclo[3.3.1]nonane without any special bond restraints. Spots in different colors represent different force constants/bond length groups of the special bond.

ring is rigid enough – and structural changes are small enough – to buffer the effect of bond length changes. However, in the bridged ring system, the geometry dictates that changes in distance between the atoms in question cannot easily be absorbed by small changes in other bond lengths, resulting in significant structural discrepancies between the conformation of the dummy atoms in water and in the binding environment, which, we expected, will lead to larger errors.

We still need some way of determining whether these effects will be significant for real binding free energy calculations, so we examined strain in several real protein-ligand binding systems. Specifically, we examined simulations of several different protein-ligand complexes and the



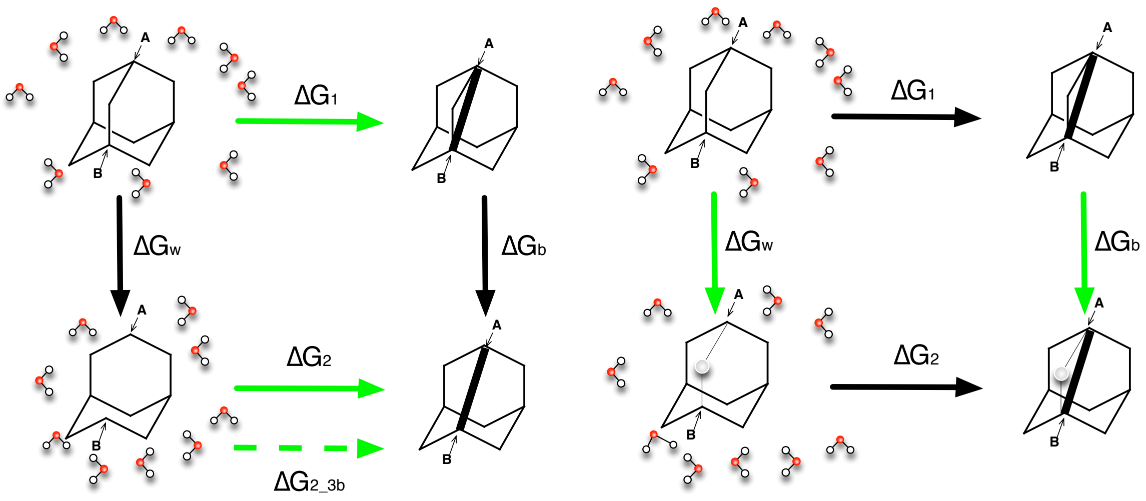
(a) absolute

(b) relative

Figure 4.6: The calculations we run are the green arrows. (a) We calculate the absolute "binding" free energy for each molecule, green arrows shown in (a), from water to the hypothetical binding site and then get the difference in absolute binding free energies as $\Delta\Delta G_{ab} = \Delta G_1 - \Delta G_2$. Since in this case $\Delta G_1 + \Delta G_b - \Delta G_2 - \Delta G_w = 0$, then $\Delta\Delta G_{ab} = \Delta G_w - \Delta G_b$. (b) We also calculate the relative "binding" free energy, green arrows shown in (b), by changing atoms to dummy atoms (hollow spheres), and get the relative binding free energy difference as $\Delta\Delta G_{rl} = \Delta G_w - \Delta G_b$. Here the bold bond on the right hand side of both panels represents the bond strained by "binding" to our hypothetical receptor (here represented by introducing restraints).

free ligands in solution to determine the magnitude of typical changes in bond length. The simulation trajectories were obtained from our former projects which include six ligands in trypsin, two ligands bound to DNA gyrase [171] as provided by Vertex Pharmaceuticals, and ibuprofen in HSA (Human serum albumin). Trajectories and parameter/coordinate files are provided in the supporting material. Our current research efforts do not provide good benchmarks for bond length changes in fused rings systems, but we believe the systems examined here provide some idea of the amount of bond length change which can be expected in general, at least enough so to give a rough idea of the size of the effect.

We found that, in most of these simulations, bond length changes were small. Bond lengths differ in the binding site by less than 1 percent from those in solution. Ibuprofen binding to HSA proved an exception – we saw somewhat larger bond length changes, with two over 1%. (Table 4). Based on work in our model systems, bond length changes of this magnitude



(a) absolute

(b) relative

Figure 4.7: The calculations we run for adamantane-bicyclononane and adamantane-bicyclononane with NDE are the green arrows ΔG_1, ΔG_b, ΔG_w with green arrow ΔG_2 for adamantane-bicyclononane and dashed green arrow ΔG_{2_3b} for adamantane-bicyclononane with NDE. The meaning of each simulation is the same as in Fig. 4.6, with panel (a) showing the absolute "binding" free energy for each molecule and panel (b) showing the relative "binding" free energy. For adamantane-bicyclononane specifically, the bottom absolute simulation ΔG_2 starts from the topology based on the structure of bicyclo[3.3.1]nonane plus topology modifications to maintain the same 1-4 and internal interactions with the end state of the left relative simulation ΔG_w. For adamantane-bicyclononane with NDE, the bottom absolute simulation ΔG_{2_3b} starts from the original topology generated from the structure of bicyclo[3.3.1]nonane.

would be sufficient to cause errors larger than 0.5 kT, which is small but notable, in the bicyclic system (Fig. 4.3) and an error as large as 1-30 kcal/mol in the cage-like system depending on how we measure the original bond length (Fig. 4.5, SI).

Data for these systems is provided in the Supporting Information.

4.5 Conclusions

Fundamentally, the error introduced by ring breaking results from coupling between dummy atoms in multiply-connected groups (such as a ring system which has been turned into dummy atoms) and the conformation of the rest of the system. Specifically, the thermodynamic cycle used for relative free energy calculations assumes that the contribution of the dummy atoms to the free energy of the system is equivalent in the different environments, which is not in general the case. Strain in the ligand or solute induces some degree of conformational change, which affects the free energy of the dummy atoms so that this assumption is no longer met.

In this study, we examined how this error introduced by ring breaking in relative free energy calculations grows as a function of ligand strain in a model binding system. We find that for bicyclic ring systems, errors are relatively small (less than 0.5 kT) and typically not statistically significant if the ligand strain is small – that is, if bond length changes caused by the binding environment are small (less than 2%). However, substantial changes in bond length as large as 1% do seem to occur in some real systems we have examined, suggesting that such perturbations ought to be avoided whenever possible. But we further find that for cage-like or bridged molecules, if we remove the bridge, errors grow much more rapidly as a function of ligand strain. In the system we examined here, even 1% distance changes lead to errors of 1 to many kcal/mol (depending on how the change in bond length is measured). Furthermore, since ligand strain is difficult to predict *a priori*, there is no way to know in advance how big these errors will be for a specific system of interest. So in all we believe ring breaking should be avoided in relative free energy calculations whenever possible, even for planar rings, though it is especially critical to avoid breaking bridged rings.

If researchers do need to calculate free energy changes for transformations involving ring breaking, we believe dual- or separated-topology[125, 146] relative free energy calculations and absolute free energy calculations[125] may be a better options, as these do not suffer from the same limitations.

4.6 Appendix

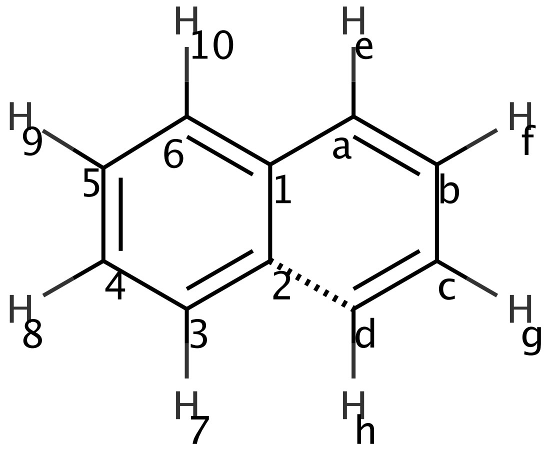


Figure 4.8: 2D structure for naphthalene. Considering a transformation from naphthalene to benzene, the dashed line highlights the fact that the (right-hand) ring which would be transformed to dummy atoms is doubly connected to the remaining ring.

In alchemical relative binding free energy calculations, the two molecules before and after mutation usually have different topologies and differing numbers of atoms, and dummy atoms are therefore necessarily introduced in the calculations. To ensure that the effect of the dummy atoms exactly cancels out in the two legs of the simulations, certain rules must be followed regarding which interaction energy terms between the dummy atoms and the physical atoms are kept in the two end points of the relative calculation – specifically, in the states reached at the end point lambda values.

In general, the end point lambda window in alchemical FEP simulations has the following parts: N physical atoms $1, 2, \dots N$ in the mixed molecule, m dummy atoms $a, b, c, \dots x$, and i atoms in the surrounding environment $S_1, S_2, \ldots S_i$ (the solvent, ion and/or protein). The total interaction energy (bonded and non-bonded) has the following components: the interaction energy between the physical atoms (U_P^{self}), the interaction energy between the dummy atoms (U_D), the interaction between dummy atoms and physical atoms (U_{PD}), the interaction energy between particles in the environment (U_e), and the interaction between the environment and the physical atoms (U_{Pe}).

$$U(1,2...N;a,b,...x;S_1,S_2,...S_i) = U_p^{self}(1,2,...N) + U_D(a,b,...x) + U_{PD}(1,2,...N;a,b,...x)+ \\ U_e(S_1,S_2,...S_i) + U_{Pe}(1,2,...N;S_1,S_2,...S_i) \tag{4.1}$$

Since the dummy atoms do not interact with the surrounding environment, we can define the following effective potential due to the surrounding environment by integrating over these degrees of freedom:

$$U_P^{env}(1,2,...N) = -kT \ln \int dS_1 dS_2 ... dS_i \exp(-\beta(U_e + U_{Pe})) \tag{4.2}$$

The effective potentials from the environments are different in the two legs of the alchemical FEP simulations. Define

$$U_P = U_P^{self} + U_P^{env} \tag{4.3}$$

Then the configurational part of the partition function for the whole system simplifies into:

$$Q = \int d1d2...dNda\ db...dx \exp(-\beta(U_P + U_D + U_{PD})) \tag{4.4}$$

It is easy to show that, if there are only one bonded stretch interaction, two bonded angle interactions, and three bonded dihedral angle interactions between the physical atoms and the dummy atoms at the end point then the effect of the dummy atoms exactly cancels out in the two simulations [12]. (These interactions could be labeled $r_{1a}, \theta_{21a}, \theta_{1ab}, \phi_{321a}, \phi_{21ab} and \phi_{1abc}$ where the subscripts stand for the atom numbers in Fig. 4.8; for example θ_{1ab} refers to the angle between bond $1a$ and bond ab.) This is also true for fewer retained interactions. The 3m degrees of freedom for the dummy atoms can be decomposed into 3m-6 internal degrees of freedom for the dummy atoms, $x_1^D, x_2^D, ...x_{3m-6}^D$, and 6 degrees of freedom joining the dummy atoms with the physical atoms $x_1^{PD}, x_2^{PD}, ...x_6^{PD}$ (the 6 degrees of freedom listed above). Therefore:

$$\begin{aligned}
&Q(1,2,...N,a,b,c,...x) = \int d1d2...dNda\ db...dx \exp(-\beta(U_P + U_D + U_{PD})) \\
&= \int d1d2...dN \exp(-\beta U_P) \int dx_1^{PD} dx_2^{PD} ...dx_6^{PD} \exp(-\beta U_{PD}) \int dx_1^D dx_2^D ...dx_{3m-6}^D \exp(-\beta U_D) \\
&= Q(1,2,...N) Q^D
\end{aligned} \tag{4.5}$$

Now suppose that, in addition to the interactions involving the 6 degrees of freedom mentioned above, one more interaction between the dummy atoms and the physical atoms is retained in the end lambda window, such as the bonded stretch interaction between atoms 2 and d in our example (Fig. 4.8) of a ring closing mutation involving perturbing a benzene ring to a napthelene ring. If this extra interaction is retained, then

$$\begin{aligned}
&Q(1,2,...N,a,b,c,...x) = \int d1d2...dNda\ db...dx \exp(-\beta(U_P + U_D + U_{PD})) \\
&= \int d1d2...dN \exp(-\beta U_P) \int dx_1^{PD} dx_2^{PD} ...dx_6^{PD} \int dx_1^D dx_2^D ...dx_{3m-6}^D \exp(-\beta(U_{PD}^6 + U_{r_{2d}})) \exp(-\beta U_D) \\
&= \int d1d2...dN \exp(-\beta(U_P + U_P^{eff}))
\end{aligned}$$

(4.6)

where $U_P^{eff}(1,2,...N) = -kT \ln \int dx_6^{PD} dx_{3m-6}^{D} \exp(-\beta(U_{PD}^6 + U_{r_{2d}} + U_D))$ is the effective potential (restraint) applied on the physical atoms due to the interactions with the dummy atoms. Note that the term $U_{r_{2d}}$ — the bond stretch potential between atom 2 and d — is introduced because of the restriction of an extra degree of freedom aside from the six rigid-body degrees of freedom. Thus, the final result in Eq. 6 cannot be separated into separate integrals as in Eq. 5 because having only those six degrees of freedom restrained is the prerequisite for ensuring the thermodynamic properties of the dummy atoms and those of the physical atoms are independent [12]. Thus, with this extra interaction, the the $U_P^{eff}(1,2,...N)$ is no longer separable into a part for the dummy atoms and a part for the interaction between physical and dummy atoms. Therefore, the inclusion of additional bonded stretch interactions for the ring opening/closing FEP calculations will introduce a conformational bias for the ligand simulated, and the effect of the dummy atoms does not cancel out in the two legs of the simulations, leading to errors in the calculations.

ID	Special bond length (Å)	Force constant($kJmol^{-1}nm^{-2}$)	Bond length in simulation(Å)	Errors($kcal/mol$)
original	none	none	1.4057(8)	0.03 ± 0.04
0	1.387	4.0033	1.3979(6)	0.19 ± 0.06
1	1.370	4.0033	1.3893(7)	0.28 ± 0.06
2	1.350	4.0033	1.3795(7)	0.39 ± 0.05
3	1.330	4.0033	1.3698(7)	0.51 ± 0.05
4	1.310	4.0033	1.3604(6)	0.64 ± 0.05
5	1.290	4.0033	1.3514(8)	0.76 ± 0.05
6	1.230	4.0033	1.3237(6)	1.25 ± 0.05
7	1.110	4.0033	1.2659(6)	2.32 ± 0.05
8	0.870	4.0033	1.1524(5)	4.72 ± 0.05
11	1.387	3.6030	1.3977(6)	0.14 ± 0.05
12	1.387	2.4020	1.4001(8)	0.16 ± 0.05
13	1.387	2.0017	1.4006(6)	0.11 ± 0.05
14	1.387	1.6013	1.4014(6)	0.07 ± 0.05
15	1.387	0.4033	1.4065(7)	0.04 ± 0.04
21	1.230	3.6030	1.3283(6)	1.24 ± 0.05
22	1.230	2.4020	1.3438(6)	0.95 ± 0.05
23	1.230	2.0017	1.3500(6)	0.78 ± 0.05
24	1.230	1.6013	1.3591(6)	0.73 ± 0.04
25	1.230	0.4033	1.3925(8)	0.30 ± 0.04

Table 4.1: Errors as a function of bond length (strain) for naphthalene-benzene.

ID	Special bond length (Å)	Force constant($kJmol^{-1}nm^{-2}$)	Bond length in simulation(Å)	Errors($kcal/mol$)
original	none	none	1.553(1)	0.04 ± 0.05
0	1.535	2.5363	1.5460(7)	0.13 ± 0.06
1	1.517	2.5363	1.5375(8)	0.06 ± 0.06
2	1.494	2.5363	1.5271(8)	0.18 ± 0.06
3	1.472	2.5363	1.5166(9)	0.32 ± 0.06
4	1.449	2.5363	1.5049(7)	0.50 ± 0.06
5	1.428	2.5363	1.4941(7)	0.50 ± 0.06
6	1.362	2.5363	1.4637(7)	0.75 ± 0.06
7	1.228	2.5363	1.3981(7)	1.24 ± 0.06
8	0.964	2.5363	1.2738(7)	2.74 ± 0.05
11	1.535	2.2827	1.5476(7)	0.09 ± 0.05
12	1.535	1.5218	1.5484(8)	-0.03 ± 0.05
13	1.535	1.2682	1.5503(8)	0.12 ± 0.05
14	1.535	1.0145	1.5512(8)	0.20 ± 0.05
15	1.535	0.2563	1.5544(9)	0.14 ± 0.05
21	1.362	2.2827	1.4628(7)	1.10 ± 0.05
22	1.362	1.5218	1.4874(8)	0.54 ± 0.05
23	1.362	1.2682	1.4949(8)	0.51 ± 0.05
24	1.362	1.0145	1.5027(9)	0.35 ± 0.05
25	1.362	0.2563	1.5373(9)	0.11 ± 0.05

Table 4.2: Errors as a function of bond length (strain) for decalin-cyclohexane.

ID	Special bond length (Å)	Force constant($kJmol^{-1}nm^{-2}$)	Distance between atom A and B in simulation(Å)	Errors($kcal/mol$)
original	none	none	3.18(7)	0.29 ± 0.03
0	2.524	2.5363	2.63(3)	-21.98 ± 0.10
1	2.494	2.5363	2.61(3)	-23.97 ± 0.10
2	2.456	2.5363	2.58(3)	-26.16 ± 0.10
3	2.421	2.5363	2.54(3)	-28.07 ± 0.10
4	2.383	2.5363	2.51(3)	-30.11 ± 0.10
5	2.347	2.5363	2.48(3)	-31.65 ± 0.10
6	2.239	2.5363	2.39(3)	-35.54 ± 0.10
7	2.019	2.5363	2.20(3)	-39.61 ± 0.10
8	1.585	2.5363	1.85(3)	-31.85 ± 0.10
11	2.524	2.2827	2.65(3)	-21.69 ± 0.08
12	2.524	1.5218	2.69(4)	-19.93 ± 0.05
13	2.524	1.2682	2.71(4)	-18.88 ± 0.05
14	2.524	1.0145	2.75(4)	-17.61 ± 0.04
15	2.524	0.2536	2.97(6)	-8.65 ± 0.03
16	2.524	0.2283	2.98(6)	-8.09 ± 0.03
17	2.524	0.1522	3.03(6)	-5.91 ± 0.03
18	2.524	0.1268	3.05(6)	-5.06 ± 0.03
19	2.524	0.1015	3.08(6)	-4.23 ± 0.03
20	2.524	0.0254	3.15(8)	-1.02 ± 0.03
21	2.239	2.2827	2.41(3)	-35.04 ± 0.08
22	2.239	1.5218	2.47(4)	-33.02 ± 0.05
23	2.239	1.2682	2.50(4)	-31.82 ± 0.05
24	2.239	1.0145	2.55(4)	-30.03 ± 0.04
25	2.239	0.2536	2.88(6)	-15.63 ± 0.03

Table 4.3: Errors as a function of bond length (strain) for adamantane-bicyclononane, using distances observed in the absolute simulation of bicyclo[3.3.1]nonane for measuring the bond length. The corresponding error figure for this table is Fig. 4.5

atom ID 1	atom ID 2	bond length in water(Å)	bond length in complex (Å)	z score	percentage
C1	C3	1.388(1)	1.394(1)	3.4	0.4
C4	C6	1.389(1)	1.399(1)	6.1	0.7
C2	C5	1.388(1)	1.394(1)	4.1	0.5
C11	C5	1.517(2)	1.524(1)	3.3	0.5
C12	C6	1.515(1)	1.528(1)	6.8	0.9
C10	C13	1.535(1)	1.544(1)	4.7	0.6
C7	O1	1.215(1)	1.213(1)	1.6	0.2
C12	C8	1.534(2)	1.546(1)	5.8	0.8
C2	C4	1.388(1)	1.393(1)	2.8	0.3
C7	O2	1.306(1)	1.291(1)	9.4	1.1
C1	C5	1.385(1)	1.394(1)	5.2	0.6
C11	C13	1.535(2)	1.552(1)	6.8	1.1
C3	C6	1.387(1)	1.398(1)	6.9	0.8
C12	C7	1.508(1)	1.521(1)	7.4	0.9
C13	C9	1.536(1)	1.549(1)	6.4	0.8

Table 4.4: Bond length changes in a real protein-ligand complex – ibuprofen binding to HSA.

Chapter 5

Using MD simulations to calculate how solvents modulate solubility

5.1 Abstract

Here, our interest is in predicting solubility in general, and we focus particularly on predicting how the solubility of particular solutes is modulated by the solvent environment. Solubility in general is extremely important, both for theoretical reasons – it provides an important probe of the balance between solute-solute and solute-solvent interactions – and for more practical reasons, such as how to control the solubility of a given solute via modulation of its environment, such as in process chemistry or separations. Here, we how the change of solvent affects the solubility of a given compound. That is, we calculate relative solubilities. We use MD simulations to calculate relative solubility and compare our calculated values with experiment as well as with results from several other methods, SMD and UNIFAC, which are more commonly used in chemical engineering. We find that straightforward solubility calculations based on molecular simulations using a general small-molecule force field outperform

GAFF and UNIFAC both in terms of accuracy and coverage of the relevant chemical space.

5.2 Introduction

Solubility is a fundamental property in industry, and is of particular interest in purification and separations. Thus, a good deal of research effort has been invested towards predicting solubility. However, in a recent blind test of current methods[73], predictions did not perform nearly as well as retrospective tests, suggesting substantial challenges remain. In part, there may be large issues of transferability of these models, which are often fairly highly parameterized based on existing data.

Several classes of methods have been employed in this area. One main category of methods is empirical methods based on molecular descriptions, like the Group Contribution(GC) method. In this category, one commonly employed method is UNIFAC [45, 61, 184, 55, 56] which uses a compound library to analyze the contribution to solubility each functional group and predict the new solubilities based that knowledge. This approach is fast, and can produce acceptable results in many cases. However, a major potential drawback of this class of methods is that GC methods require a good deal of experimental data to calculate the contributions of each functional group. If a functional group does not exist in the experimental library, then solubility predictions for compounds with this functional group cannot be expected to be accurate. A second category includes statistical methods like multiple linear regression (MLR) or Neural Network (NN) methods [70]. These methods use statistical or machine-learning tools to analyze the existing data, build a model, polish the parameters of the model, test the model and then use the created model to predict solubility. Some of these methods have good results [75, 175, 143], with RMS errors (RMSE) around 1.0 log unit and correlation coefficients (R^2) around 0.8. However, these models require a large amount of high quality input data for training, which can pose challenges. Additionally, the physical interpretation of each model can pose challenges. Overall, both major classes of method can easily suffer from problems of transferability, as illustrated by the recent blind tests [73]. This is likely because these methods are highly dependent on the size and quality of the training set, and because of the degree of human input required in building the models.

Aside from these empirical methods, there have been relatively few simulation-based efforts to calculate solubilities or relative solubilities from physical principles rather than empirical training [134, 2, 19]. Here, we will call calculations based on physical principles "direct" solubility calculations, and in our view direct calculations are those which do not require training on solubility data, and do not require human interpretation of or adjustment of the model. Rather, direct calculations typically involve calculation of the underlying thermodynamic contributions to solubility (the chemical potentials of the solute in solid versus in solution) or approximations thereof.

Here, we focus on using simulations to calculate solubilities, but focus particularly on relative solubilities. It is still relatively difficult to compute the excess chemical potential of the solid or related properties as needed for absolute solubility calculations[44]. Focusing on calculating the relative solubility of a solute in different solvents allows us to focus on solution-phase thermodynamics of the solute and how these are affected by the solvent. In other words, we can still directly calculate *relative* solubilities of the same solute in different solvents even without information about the chemical potential or free energy of the solid. Details of our approach can be found below in Methods. Here, we compute solubilities for eight solutes in 34 different solvents, for a total of 53 different solute-solvent pairs. Data for our test comes from the Open Notebook Science Solubility Challenge [15]. For each of these solute-solvent pairs, we compute the solvation free energy and other properties, allowing us to calculate the relative solubility for comparison with experiment. We also compare our methods with two other commonly used methods UNIFAC[45, 61, 184, 55, 56] and SMD[109, 108, 145] and find that our calculations are more accurate than those from the stated methods on the present set, and also cover more of the compounds in our set.

5.3 Methods

5.3.1 Theory

To calculate the solubility of a single solute in a particular solvent directly, we need to know two pieces of information, the solvation free energy, and the fugacity of pure solid solute. Given these, the solubility can be calculated via:

$$\ln x_1^\alpha = -\beta {\mu_1^{\alpha,res}}(T,p,x_1) - \ln(\frac{RT}{v(T,p,x_1)}) + \ln f_1^S(T,p) \tag{5.1}$$

where x_1^α is the equilibrium solubility of the solute in units of mole fraction, $\beta {\mu_1^{\alpha,res}}(T,p,x_1)$ is the dimensionless residual chemical potential of the solute (denoted by the subscript 1) in solvent α, v is the molar volume of the mixture (solvent 1 plus solvent α), and f_1^S is the fugacity of pure solid solute.

In concentration units (molar), this could be rewritten as:

$$\ln c_1^\alpha = \ln(\frac{x_1^\alpha}{v(T,p,x_1)}) - \beta {\mu_1^{\alpha,res}}(T,p,x_1) - \ln(RT) + \ln f_1^S(T,p) \tag{5.2}$$

where c_1^α is the molar concentration (at the solubility limit) of solute 1 in solvent α.

From equation 2, since f_1^S is a solute dependent constant and RT is constant, we can compute the relative solubility of the solute 1 in solvent α relative to solvent ζ as

$$\ln(\frac{c_1^\alpha}{c_1^\zeta}) = \ln(\frac{x_1^\alpha}{x_1^\zeta}\frac{v^\zeta(T,p,x_1^\zeta)}{v^\alpha(T,p,x_1^\alpha)}) = \beta\mu_1^{\zeta,res}(T,p,x_1) - \beta\mu_1^{\alpha,res}(T,p,x_1) \tag{5.3}$$

where here v^α and v^ζ correspond to the molar volume of the binary mixture of the solute in solvent α and ζ, respectively.

If we assume that the solute is infinitely dilute, then the molar volume is independent of the solute concentration or mole fraction, and so:

$$\ln(\frac{c_1^\alpha}{c_1^\zeta}) = \ln(\frac{x_1^\alpha}{x_1^\zeta}\frac{v_\zeta(T,p)}{v_\alpha(T,p)}) = \beta\mu_1^{\zeta,res,\infty}(T,p) - \beta\mu_1^{\alpha,res,\infty}(T,p) \tag{5.4}$$

where the molar volumes are now for the pure solvent, and the excess chemical potential is at infinite dilution (superscript ∞).

In this case, the residual chemical potential is equal to the Gibbs free energy of solvation of a single solvent molecule:

$$\mu_1^{\alpha,res,\infty} = \Delta G_{1,solv}^{\alpha,\infty} \tag{5.5}$$

So equation 5.5 allows us to calculate experimental relative solubilities (on the left hand side) from solvation free energies and other properties we can easily obtain from molecular simulations (right hand side). Equation 5.5 is a relative formula, comparing the solubility of the same solute in different solvents. Thus, we can compute solvation free energies for a single solute in different solvents and calculate a relative solubility for direct comparison with experiment. This approach can be used even in the absence of knowledge of the experimental crystal structure of the solid, which can be difficult to calculate [refs], and its fugacity ($\ln f_1^S(T,p)$ in equation 1), which can be even more difficult to calculate.

5.3.2 Dataset selection

To compare calculated solubilities, we drew on the Open Notebook Science Solubility Challenge [15] which provides 9700 experimental solubility datasets, where in their terminology, a "dataset" consists of a set of experimental data resulting in a solubility measurement. We wanted a test set consisting of around 50 solubility measurements, so we filtered these 9700 measurements to select a sub-set based on four rules. First, we picked cases where the number of solute heavy atoms was less than 15. Second, we focused on molecules only containing carbon, hydrogen, nitrogen, and oxygen. Third, we focused on molecules with a formal charge of zero. And fourth, we limited the number of rotatable bonds to three or less. While none of these rules represent fundamental limits of the methods we employ here, they do allow us to focus on a subset of available data, and specifically on cases where we expect conformational sampling to be relatively straightforward and force field issues to be fairly well understood. Additionally, challenges relating to the calculation of solvation free energies of charged species[87, 88] can be avoided. We also required that the concentration should be under 0.1 mole fraction to meet our infinite dilution assumption as given in Eq. 5.4. This still left us with more solute-solvent pairs than needed, so we manually selected the final set, ensuring that each solute appears at least twice (to be able to calculate the relative solubility); that a wide range of topologies are considered (including chains, rings (both aromatic and not), and polycyclic rings). We also deliberately avoided most carboxylic acids, as these could undergo a change of protonation state on transfer between different solvents, though we included two such molecules as a test. Our final set consists of 53 solute-solvent pairs, as detailed in Table 1.

5.3.3 Simulation

Our approach here is to use alchemical free energy calculations based on molecular dynamics simulations to compute solvation free energies for solutes in solution. The underlying molecular dynamics simulations can also be analyzed to calculate the molar volume.

After construction of our test set, we consider all solute-solvent pairs and generate input files for free energy calculations. For each solute or solvent, we take the SMILES string and generate 3D structures using OpenEye OEChem Python toolkit and Omega [66], then assign AM1-BCC [78, 79] partial charges. Antechamber [178] from AmberTools 13 was used to assign GAFF [179] atom types and then AmberTools' tleap was used to generate assign GAFF parameters [179] and write AMBER .prmtop and .crd files. The resulting files were converted to GROMACS format using acpype [160]. The individual solute and solvent GROMACS input files were stored, and we then used packmol [110] to create solvated boxes consisting of one solute surrounded by many different solvent molecules. The boxes were cubical simulation, with at least 1.2 nm from the solute to the nearest box edge.

AMBER combination rules (arithmetic average for σ and geometric for ε) were used. Simulations were run using Langevin dynamics, as previously [93], and the simulation timestep was 1 fs. Lennard-Jones interactions were gradually switched off between 0.9 and 1.0 nm, and an analytical correction was applied to the energy and pressure. PME was used for electrostatics, as previously, with a real-space cutoff of 1.2 nm. LINCS was used to constrain bonds to hydrogen. Each system and λ value (where λ is a parameter ranging between 0 and 1, where 0 corresponds to the unmodified system, and 1 corresponds to the end-state of the transformation) was independently minimized for up to 2500 steps of steepest-descents

Solute ID in Solvent ID	Solute Name	Solvent Name
10241 in 10907	9-fluorenone	2,2,4-trimethylpentane
10241 in 3283	9-fluorenone	diethyl ether
10241 in 6276	9-fluorenone	1-pentanol
10241 in 6342	9-fluorenone	acetonitrile
10241 in 7298	9-fluorenone	cyclopentanol
10241 in 8058	9-fluorenone	n-hexane
243 in 1140	benzoic acid	toluene
243 in 174	benzoic acid	ethylene glycol
243 in 241	benzoic acid	benzene
243 in 6342	benzoic acid	acetonitrile
243 in 8003	benzoic acid	pentane
243 in 8058	benzoic acid	n-hexane
243 in 8078	benzoic acid	cyclohexane
243 in 887	benzoic acid	methanol
2519 in 180	caffiene	acetone
2519 in 887	caffiene	methanol
2519 in 962	caffiene	water
638088 in 1031	trans-stilbene	1-propanol
638088 in 10907	trans-stilbene	2,2,4-trimethylpentane
638088 in 1140	trans-stilbene	toluene
638088 in 18508	trans-stilbene	tert-butylcyclohexane
638088 in 241	trans-stilbene	benzene
638088 in 263	trans-stilbene	1-butanol
638088 in 31275	trans-stilbene	1,4-dioxane
638088 in 3776	trans-stilbene	2-propanol
638088 in 6276	trans-stilbene	n-pentanol
638088 in 6560	trans-stilbene	isobutyl alcohol
638088 in 702	trans-stilbene	ethanol
638088 in 7929	trans-stilbene	3-xylene
638088 in 8028	trans-stilbene	tetrahydrofuran
638088 in 8058	trans-stilbene	n-hexane
638088 in 887	trans-stilbene	methanol
7107 in 3283	xanthene	diethyl ether
7107 in 702	xanthene	ethanol
7107 in 7914	xanthene	isopropyl ether
7107 in 8078	xanthene	cyclohexane
7107 in 887	xanthene	methanol
7478 in 31275	4-methoxybenzoic acid	1,4-dioxane
7478 in 6276	4-methoxybenzoic acid	n-pentanol
7478 in 6560	4-methoxybenzoic acid	isobutyl alcohol
7478 in 8028	4-methoxybenzoic acid	tetrahydrofuran
77577 in 180	2,3-dimethyl-2,3-dinitrobutane	acetone
77577 in 31275	2,3-dimethyl-2,3-dinitrobutane	1,4-dioxane
77577 in 6342	2,3-dimethyl-2,3-dinitrobutane	acetonitrile
77577 in 6569	2,3-dimethyl-2,3-dinitrobutane	methylethyl ketone
77577 in 7967	2,3-dimethyl-2,3-dinitrobutane	cyclohexane
77577 in 8028	2,3-dimethyl-2,3-dinitrobutane	tetrohydrofuran
77577 in 8857	2,3-dimethyl-2,3-dinitrobutane	ethyl acetate
8418 in 6228	anthracene	dimethylformamide
8418 in 7237	anthracene	2-xylene
8418 in 7505	anthracene	benzonitrile
8418 in 7929	anthracene	3-xylene

Table 5.1: Solute-solvent pairs used for simulation. Here, we use PubChem compound identifiers to track our compounds as these are easier for our tools to work with than full chemical names; the table lists both.

minimization.

Following constant pressure equilibration, box sizes were adjusted at each λ value by an affine transformation to ensure each λ value had the correct volume for the target pressure. After this, we conducted an additional 5 ns of constant pressure production simulation at each λ, discarding the first 100 ps as additional "equilibration", as previously [93]. Here we use Parrinello-Rahman barostat to modulate the pressure.

The parameter λ controls the transformation between end states. In this version of GROMACS, we use two separate λ values, one controlling modification of partial charges (λ_{chg}, turning solute partial charges to zero), the second controlling modification of Lennard-Jones interactions (λ_{LJ}, turning solute LJ interactions to zero). Here, λ_{chg} = [0.0 0.25 0.5 0.75 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0] and λ_{LJ} = [0.0 0.00 0.0 0.00 0.0 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1.0]. Hence, partial charges are first turned off and then LJ interactions are turned off separately. Hydration free energies are calculated by first computing the free energy of turning of solute-environment interactions in gas (vacuum) and then subtracting the free energy of turning these off in water.

Additionally, the Mobley lab recently developed new GAFF-DC hydroxyl parameters [43], a modification of the original GAFF parameter set. To test these parameters, all hydroxyl-containing molecules were run both with the original GAFF parameter set and with GAFF-DC.

5.3.4 Other methods

Beside the simulation methods, we also use the SMD and UNIFAC methods which are commonly used in chemical engineering In total, we used six methods to calculate relative solubilities, which we label as follows:

1. GAFF: Alchemical free energy calculations with standard GAFF
2. GAFF-DC: Alchemical free energy calculations with GAFF-DC [43]
3. SMD: SMD using the solvent optimized geometry
4. SMD_vac: SMD using the original vacuum optimized geometry [109, 108, 145]
5. UNIFAC: The UNIFAC approach [61, 184]
6. UNIFAC_mod: A slightly modified the functional formal of UNIFAC [55, 56]

Results from these approaches will be discussed below.

5.4 Results

Our free energy calculations allow us to calculate the different terms on the right side of Eq. 5.3 – that is, the difference in excess chemical potentials μ, or the difference in solvation free energies for a particular solute in a particular pair of solvents. We call this value the calculated value. We can then directly compare with the experimental relative solubility – the term involving $\ln c$ on the left side of Eq. 5.3. We call this the experimental value. The error for a particular solute-solvent pair is then taken as the difference between the calculated value and the experimental value.

Analysis is made slightly more complicated by the fact that we can actually calculate many different errors which are interrelated. So for one solute, if there are n different solvents, there are $n * (n-1)/2$ possible solvent pairs we could examine, leading to $n * (n-1)/2$ potential errors. We call the set of all possible errors for a specific solute (for all solvent pairs we examine that solute in) a ‘dataset’ and we use the PubChem Compound Identifier (CID) of the solute as the title of this dataset.

UNIFAC and SMD errors were analyzed via the same approach.

5.1 shows the average of the all these pairwise errors for each solute, for each of the methods examined here. Tables 2—7 show error statistics for these methods. The all pairwise results suggested that the simulation with new hydroxyl parameter get the lowest errors among all methods.

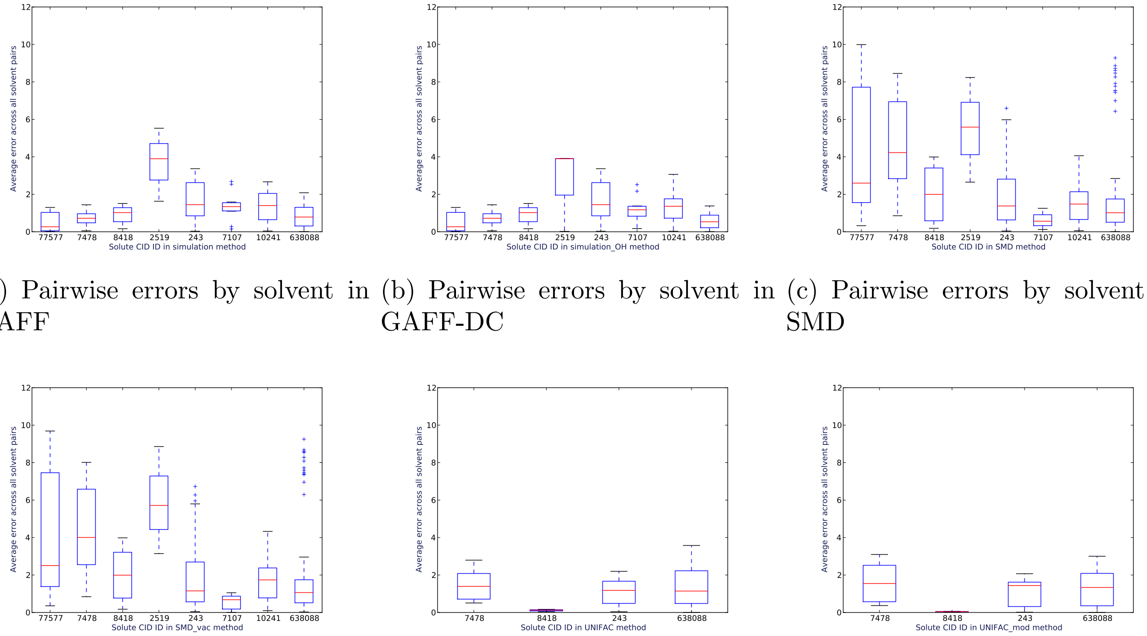
We are also interested in understanding not just the error in our calculated values, but how well they reproduce experimental values. Thus, we plot experimental relative solubilities versus calculated ones - specifically, experimental vs calculated $\ln(\frac{c_1^\alpha}{c_1^\zeta})$ – in 5.2. We find that our approach based on full free energy calculations with GAFF or GAFF-DC performs best in terms of correlation (R^2) with experiment. In contrast, SMD yields very low correlation with experiment. UNIFAC, in contrast, tends to have fairly small errors - but the R^2 is smaller than the alchemical GAFF-based approaches. Additionally, for both SMD and UNIFAC (especially UNIFAC) compound coverage is not as good, so the size of the analyzed dataset is smaller. These techniques simply do not cover all solute-solvent combinations examined here (5.1 and Tables 2—7) because of their need for training data.

Additionally, we seek to compare performance across different methods directly, so we plotted errors on each solute (across all solvents) for different methods. Specifically, 5.3 shows the error for each solute from our standard alchemical GAFF approach on the horizontal axis, versus the error on the same compounds with an alternate approach on the vertical axis. If both methods performed equally well or equally poorly, all data points would fall on the blue $x = y$ line. On the other hand, whenever the method showing on the vertical axis performs better than that on the horizontal axis, the data point will fall below $x = y$ (between $x = y$ and the x axis), and vise versa. In general we find far more points above the line than below, indicating that our GAFF approach typically outperforms the other approaches studied, except GAFF-DC.

Another way to examine our results is to actually use the experimental solubility for a solute in one or more specific solvents to determine an estimate of the fugacity term in Equation 5.2, then compare that to the estimates of the fugacity term which we would have obtained if we had done the same with other solvents. The downside of this, however, is that we have to pick one or more particular experimental solubilities to use to estimate the fugacity term. But this approach also allows us to examine whether the average error for a particular compound across all solvents might appear unusually large simply because of a large error for just one individual solvent. To investigate this, for each compound we selected one solvent to use as a test case, and used the remaining solvents as a "training set" to determine the fugacity term in Equation 5.2. A schematic of this is shown in 5.4. In this example, solute A is solvated in solvent B, C and D. So first we pick solvent B as the our test case, and use solvents C and D as the training solvents to determine the fugacity term. We then estimate the fugacity term as $\ln f_{ave} = \text{mean}(\ln f_C + \ln f_D)$, where f_C and f_D are the fugacities as estimated from finding f_C such that

$$\ln c^C_{A,\text{expt}} = \ln c^C_{A,\text{calc}} = \ln(\frac{x^C_A}{v(T,p,x_C)}) - \beta\mu^{\alpha,res}_C(T,p,x_C) - \ln(RT) + \ln f^S_C(T,p) \quad (5.6)$$

where $c^C_{A,\text{expt}}$ is the experimental solubility for A in C, and $c^C_{A,\text{calc}}$ is the calculated solubility for A in C. We do the same to obtain f_D. We then calculate the error in the fugacity for our test solvent as $\delta \ln f_\alpha = \ln f_{\text{ave}} - \ln f_\alpha$ (where α denotes the selected solvent), so for example for solvent B, $\delta \ln f_B = \ln f_{\text{ave}} - \ln f_B$. This is a fair test, since B was not included when obtaining $\ln f_{\text{ave}}$. We can also calculate $\delta \ln f_C$ and $\delta \ln f_D$, though these will obviously underestimate of the true error in the fugacity since solvents C and D were included in obtaining $\ln f_{ave}$. Still, we can determine the average or RMS error (RMSE) for compounds in the "training set". In this case, the RMSE on the training set is the RMS error across $\delta \ln f_C$ and $\delta \ln f_D$. We define the "training set error" as this RMSE. This whole process of examining a particular solute, picking a particular solvent as a test case, and evaluating training set and test set errors, can be iterated across all choices of solvent. In our example of three solvents, each of B, C, and D serve as the test case in turn. This allows us to obtain three different estimates of the test set error, and three estimates of the RMSE on the training set. We take the RMS error across our test cases as the final error estimate for this particular solute dataset, and the average of the training set RMS errors as the training set error. These are shown in Tables 2—7 and suggest that the GAFF-based alchemical results are the most accurate overall.



(a) Pairwise errors by solvent in GAFF

(b) Pairwise errors by solvent in GAFF-DC

(c) Pairwise errors by solvent in SMD

(d) Pairwise errors by solvent in SMD_vac

(e) Pairwise errors by solvent in UNIFAC

(f) Pairwise errors by solvent in UNIFAC_mod

Figure 5.1: The average error in $\ln(\frac{c_1^\alpha}{c_1^\zeta})$ by solvent, across all pairs of solutes for the different methods considered (a-f). The vertical axis is unitless, and the horizontal axis shows the solvent considered. The plot is a box and whisker plot, with the box showing the lower and upper quartiles of the data, and the red line marking the median. The whiskers show the range of the data.

Solute ID	data size	all pair error	all pair absolute error	test set error	training set error
77577	7	−0.057(1)	0.561(1)	0.5(1)	0.6(4)
7478	4	−0.3859(9)	0.7289(9)	0.47(3)	0.7(1)
8418	4	−0.3284(6)	0.9130(6)	0.59(3)	0.8(1)
2519	3	−2.5980(5)	3.6839(5)	1.84(9)	3.5(3)
243	8	−0.1839(2)	1.6187(2)	1.26(6)	1.5(2)
7107	5	−0.8687(1)	1.3623(1)	0.95(5)	1.2(2)
10241	7	0.65786(8)	1.29910(8)	0.98(5)	1.2(2)
638088	15	0.20307(2)	0.83181(2)	0.68(4)	0.7(2)
Average		0.04012(4)	1.03129(4)	0.852(2)	1.089(8)

Table 5.2: GAFF error

Solute ID	data size	all pair error	all pair absolute error	test set error	training set error
77577	7	−0.057(1)	0.561(1)	0.5(1)	0.6(4)
7478	4	−0.3859(9)	0.7289(9)	0.47(3)	0.7(1)
8418	4	−0.3284(6)	0.9130(6)	0.59(3)	0.8(1)
2519	3	−2.5980(5)	2.6114(5)	1.31(7)	2.8(3)
243	8	−0.1839(2)	1.6187(2)	1.26(6)	1.5(2)
7107	5	−0.8146(1)	1.2005(1)	0.83(4)	1.1(1)
10241	7	0.76278(8)	1.40036(8)	1.05(9)	1.2(3)
638088	15	0.18366(2)	0.53873(2)	0.44(3)	0.5(1)
Average		0.04366(4)	0.86389(4)	0.751(2)	0.972(9)

Table 5.3: GAFF-DC errors

Solute ID	data size	all pair error	all pair absolute error	test set error	training set error
77577	7	1.036(0)	4.499(0)	3.616(0)	4.303(0)
7478	4	−2.721(0)	4.665(0)	2.975(0)	4.393(0)
8418	4	1.605(0)	2.027(0)	1.350(0)	2.092(0)
2519	3	5.491(0)	5.491(0)	2.746(0)	5.150(0)
243	8	0.763(0)	2.136(0)	1.852(0)	2.186(0)
7107	5	−0.411(0)	0.624(0)	0.438(0)	0.575(0)
10241	7	−0.452(0)	1.702(0)	1.318(0)	1.571(0)
638088	14	−1.141(0)	1.958(0)	2.149(0)	2.357(0)
Average		−0.347(0)	2.318(0)	2.061(0)	2.613(0)

Table 5.4: SMD errors

Solute ID	data size	all pair error	all pair absolute error	test set error	training set error
77577	7	0.944(0)	4.321(0)	3.475(0)	4.136(0)
7478	4	−2.506(0)	4.396(0)	2.805(0)	4.153(0)
8418	4	1.480(0)	2.022(0)	1.332(0)	2.038(0)
2519	3	5.905(0)	5.905(0)	2.952(0)	5.499(0)
243	8	0.789(0)	2.113(0)	1.905(0)	2.261(0)
7107	5	−0.365(0)	0.572(0)	0.415(0)	0.540(0)
10241	7	−0.506(0)	1.840(0)	1.416(0)	1.686(0)
638088	14	−1.137(0)	1.933(0)	2.120(0)	2.325(0)
Average		−0.346(0)	2.293(0)	2.051(0)	2.603(0)

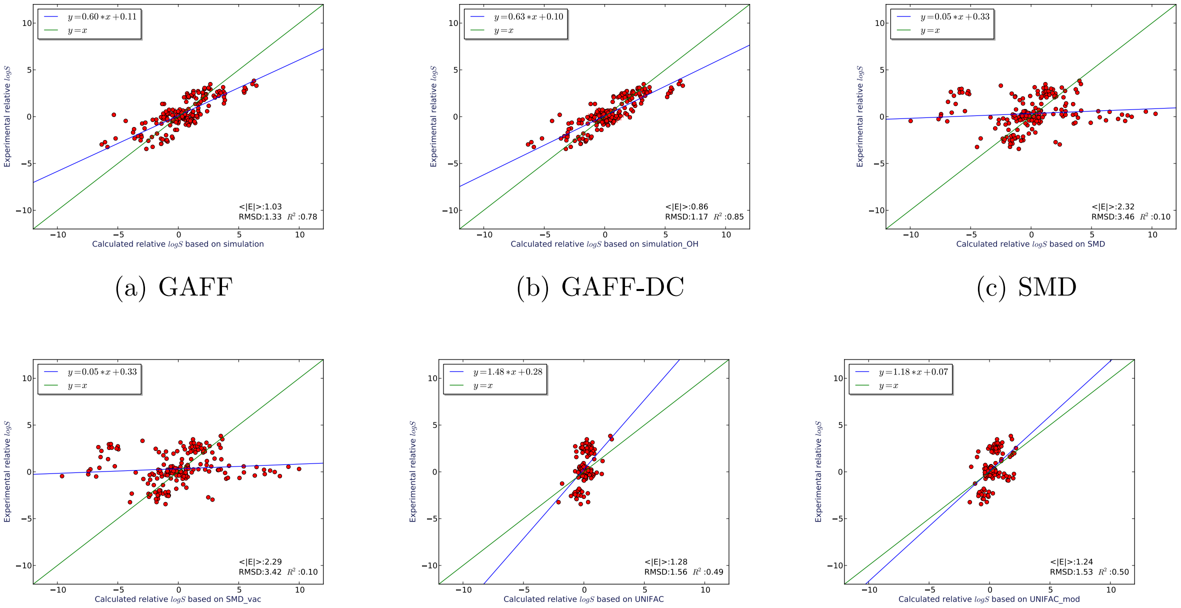
Table 5.5: SMD vac errors

Solute ID	data size	all pair error	all pair absolute error	test set error	training set error
7478	4	−0.915(0)	1.481(0)	0.936(0)	1.389(0)
8418	3	−0.028(0)	0.113(0)	0.056(0)	0.108(0)
243	6	−0.280(0)	1.123(0)	0.831(0)	1.020(0)
638088	15	−0.298(0)	1.323(0)	1.117(0)	1.201(0)
Average		−0.319(0)	1.279(0)	0.916(0)	1.072(0)

Table 5.6: UNIFAC errors

Solute ID	data size	all pair error	all pair absolute error	test set error	training set error
7478	4	−0.861(0)	1.609(0)	1.042(0)	1.583(0)
8418	3	−0.028(0)	0.042(0)	0.021(0)	0.039(0)
243	6	−0.289(0)	1.099(0)	0.840(0)	1.030(0)
638088	15	−0.061(0)	1.269(0)	1.065(0)	1.145(0)
Average		−0.124(0)	1.236(0)	0.902(0)	1.064(0)

Table 5.7: UNIFAC mod errors



(a) GAFF

(b) GAFF-DC

(c) SMD

(d) SMD_vac

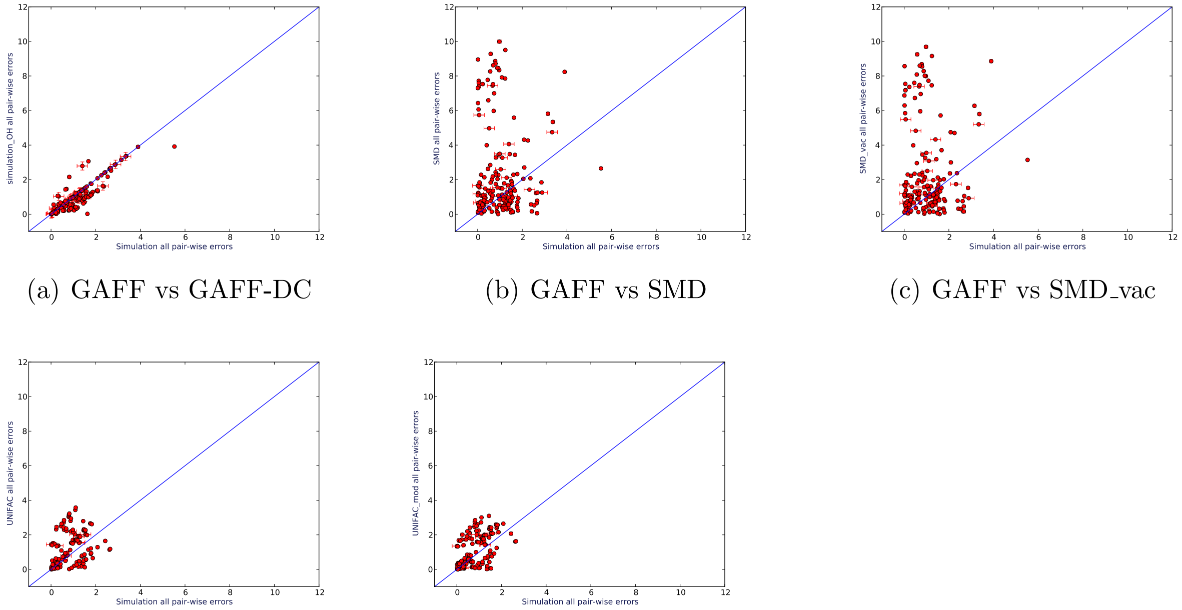
(e) UNIFAC

(f) UNIFAC_mod

Figure 5.2: Comparison of the calculated solubility with the experimental solubility for all solute solvent pairs and all methods.

5.5 Conclusions

We used alchemical free energy calculations based on molecular simulations to calculate the relative solubilities of particular solutes solvated in a variety of different solvents, achieving average absolute errors of about 1 log unit in relative solubility. We also compared our results with those obtained from SMD and UNIFAC solvation models applied to the essentially the same set, and found that our alchemical approach is more accurate in calculating relative solubilities on this set, especially when using the new GAFF-DC parameters for hydroxyl-containing compounds. Additionally, GAFF with alchemical techniques at present covers a wider range of chemical space than SMD and UNIFAC, in part because of the empirical tuning these techniques have required. We also found that overall, the GAFF-DC parameters outperform standard GAFF parameters for relative solubilities in this set. It is interesting to note that relative solubility calculations - which essentially amount to calculating a difference in solvation free energies - may be a valuable source of experimental solvation data which



(a) GAFF vs GAFF-DC (b) GAFF vs SMD (c) GAFF vs SMD_vac

(d) GAFF vs UNIFAC (e) GAFF vs UNIFAC_mod

Figure 5.3: Comparison of the all pairwise errors of the simulation method with all other methods.

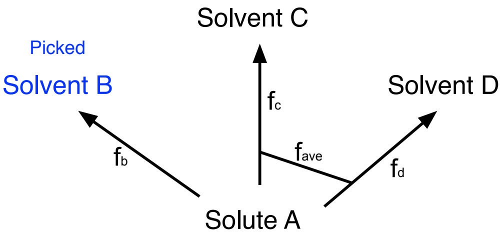


Figure 5.4: Example of how we calculate the test and training set errors.

can perhaps be used to further test and improve force fields for molecular simulations.

Bibliography

[1] M. E. Abram, R. M. Hluhanich, D. D. Goodman, K. N. Andreatta, N. A. Margot, L. Ye, A. Niedziela-Majka, T. L. Barnes, N. Novikov, X. Chen, E. S. Svarovskaia, D. J. McColl, K. L. White, and M. D. Miller. Impact of Primary Elvitegravir Resistance-Associated Mutations in HIV-1 Integrase on Drug Susceptibility and Viral Replication Fitness. *Antimicrobial Agents and Chemotherapy*, 57(6):2654–2663, June 2013.

[2] B. Aguilar and A. V. Onufriev. Efficient Computation of the Total Solvation Energy of Small Molecules via the R6 Generalized Born Model. *J. Chem. Theory Comput.*, 8(7):2404–2411, July 2012.

[3] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 1993.

[4] M. L. Benson, J. C. Faver, M. N. Ucisik, D. S. Dashti, Z. Zheng, and K. M. Merz, Jr. Prediction of trypsin/molecular fragment binding affinities by free energy decomposition and empirical scores. *J Comput Aided Mol Des*, 26(5):647–659, Apr. 2012.

[5] T. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539, June 1994.

[6] D. M. Blow. Rearrangement of Cruickshank's formulae for the diffraction-component precision index. *Acta Crystallographica Section D: Biological Crystallography*, 58(5):792–797, Apr. 2002.

[7] A. Bondy and U. Murty. *Graph Theory*. Graduate Texts in Mathematics. Springer, 2008.

[8] S. Boresch and S. Bruckner. Avoiding the van der Waals endpoint problem using serial atomic insertion. *Journal of Computational Chemistry*, 32(11):2449–2458, Aug. 2011.

[9] S. Boresch and M. Karplus. The role of bonded terms in free energy simulations. 2. calculation of their influence on free energy differences of solvation. *The Journal of Physical Chemistry A*, 103(1):119–136, 1999.

[10] S. Boresch and M. Karplus. The role of bonded terms in free energy simulations: 1. theoretical analysis. *The Journal of Physical Chemistry A*, 103(1):103–118, 1999.

[11] S. Boresch and M. Karplus. The Role of Bonded Terms in Free Energy Simulations. 2. Calculation of Their Influence on Free Energy Differences of Solvation. *J. Phys. Chem. A*, 103(1):119–136, 1999.

[12] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *The Journal of Physical Chemistry B*, 107:9535–9551, 2003.

[13] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossváry, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, page 84, Tampa, FL, 2006.

[14] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill, and B. K. Shoichet. Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *Journal of Molecular Biology*, 394(4):747–763, Dec. 2009.

[15] J.-C. Bradley, B. Friesen, J. Mancinelli, T. Bohinski, K. Mirza, D. Bulger, M. Moritz, M. Federici, D. Rein, C. Tchakounte, J.-C. Bradley, H. Truong, C. Neylon, R. Guha, A. Williams, B. Hooker, J. Hale, and A. Lang. Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents. *Nature Precedings*, Mar. 2010.

[16] E. Cancès, B. Mennucci, and J. Tomasi. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *The Journal of Chemical Physics*, 107(8):3032, 1997.

[17] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, Dec. 2005.

[18] D. S. Cerutti, N. A. Baker, and J. A. McCammon. Solvent reaction field potential inside an uncharged globular protein: A bridge between implicit and explicit solvent models? *The Journal of Chemical Physics*, 127(15):155101, 2007.

[19] L. Chebil, C. Chipot, F. Archambault, C. Humeau, J. M. Engasser, M. Ghoul, and F. Dehez. Solubilities Inferred from the Combination of Experiment and Simulation. Case Study of Quercetin in a Variety of Solvents. *The Journal of Physical Chemistry B*, 114(38):12308–12313, Sept. 2010.

[20] P. Cherepanov, A. L. B. Ambrosio, S. Rahman, T. Ellenberger, and A. Engelman. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proc. Nat. Acad. Sci. USA*, 102(48):17308–17313, Nov. 2005.

[21] C. Chipot. Free Energy Calculations in Biological Systems. How Useful Are They in Practice? *New algorithms for macromolecular simulation*, 2006.

[22] C. Chipot, X. Rozanska, and S. B. Dixit. Can free energy calculations be fast and accurate at the same time? Binding of low-affinity, non-peptide inhibitors to the SH2

domain of the src protein. *Journal of Computer-Aided Molecular Design*, 19(11):765–770, Dec. 2005.

[23] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande. Alchemical free energy methods for drug discovery: progress and challenges. *Current Opinion in Structural Biology*, 21(2):150–160, Feb. 2011.

[24] C. D. Christ, A. E. Mark, and W. van Gunsteren. Basic ingredients of free energy calculations: A review. *Journal of Computational Chemistry*, 31(8):1569–1582, 2010.

[25] F. Christ, S. Shaw, J. Demeulemeester, B. A. Desimmie, A. Marchand, S. Butler, W. Smets, P. Chaltin, M. Westby, Z. Debyser, and C. Pickford. Small-Molecule Inhibitors of the LEDGF/p75 Binding Site of Integrase Block HIV Replication and Modulate Integrase Multimerization. *Antimicrobial Agents and Chemotherapy*, 56(8):4365–4374, Aug. 2012.

[26] F. Christ, A. Voet, A. Marchand, S. Nicolet, B. A. Desimmie, D. Marchand, D. Bardiot, N. J. Van der Veken, B. Van Remoortel, S. V. Strelkov, M. De Maeyer, P. Chaltin, and Z. Debyser. Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. *Nat Meth*, 6(6):442–448, May 2010.

[27] R. G. Coleman, T. Sterling, and D. R. Weiss. SAMPL4 & DOCK3.7: Lessons for automated docking procedures. *J Comput Aided Mol Des*, 2014.

[28] C. R. Corbeil, T. Sulea, and E. O. Purisima. Rapid Prediction of Solvation Free Energy. 2. The First-Shell Hydration (FiSH) Continuum Model. *J Chem Theory Comput.*, 6(5):1622–1637, May 2010.

[29] E. De Clercq. Perspectives of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. *Il Farmaco*, 54(1–2):26–45, Jan. 1999.

[30] C. de Graaf, C. Oostenbrink, P. H. J. Keizers, B. M. A. van Vugt-Lussenburg, J. N. M. Commandeur, and N. P. E. Vermeulen. Molecular Modeling-Guided Site-Directed Mutagenesis of Cytochrome P450 2D6. *Current Drug Metabolism*, 8(1):59–77, 2007.

[31] A. de Ruiter and C. Oostenbrink. Efficient and accurate free energy calculations on trypsin inhibitors. *Journal of Chemical Theory and Computation*, 8(10):3686–3695, 2012.

[32] Y. Deng and B. Roux. Hydration of amino acid side chains: nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules. *The Journal of Physical Chemistry B*, 108:16567–16576, 2004.

[33] Y. Deng and B. Roux. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *Journal of Chemical Theory and Computation*, 2(5):1255–1273, Sept. 2006.

[34] Y. Deng and B. Roux. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *The Journal of Physical Chemistry B*, 113(8):2234–2246, 2009.

[35] T. G. Dewdney, Y. Wang, I. A. Kovari, S. J. Reiter, and L. C. Kovari. Reduced HIV-1 integrase flexibility as a mechanism for raltegravir resistance. *Journal of Structural Biology*, 184:245–250, 2013.

[36] R. Dixon. OEAntechamber: Assign and generate AMBER atom types and structural parameters. *SimTk.org – OEAntechamber: Assign and generate AMBER atom types and structural parameters*, Mar. 2011.

[37] J. Dolenc, C. Oostenbrink, J. Koller, and W. van Gunsteren. Molecular dynamics simulations and free energy calculations of netropsin and distamycin binding to an AAAAA DNA binding site. *Nucleic acids research*, 33(2):725, 2005.

[38] J. Ellson, E. Gansner, E. Koutsofios, S. North, and G. Woodhull. Graphviz and dynagraph – static and dynamic graph drawing tools. In M. Junger and P. Mutzel, editors, *Graph Drawing Software*, pages 127–148. Springer-Verlag, 2003.

[39] A. Engelman and P. Cherepanov. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat Rev Micro*, 10(4):279–290, 2012.

[40] I. J. Enyedy and W. J. Egan. Can we use docking and scoring for hit-to-lead optimization? *Journal of Computer-Aided Molecular Design*, 22:161–168, 2008.

[41] P. W. Erhardt and J. R. Proudfoot. Drug discovery: historical perspective, current status, and outlook. *Compr Med Chem II*, 1:29–96, 2007.

[42] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des*, 15(5):411–428, 2001.

[43] C. J. Fennell, K. L. Wymer, and D. L. Mobley. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. *The Journal of Physical Chemistry B*, 118(24):6438–6446, June 2014.

[44] M. Ferrario, G. Ciccotti, E. Spohr, and T. Cartailler. Solubility of KF in water by molecular dynamics using the Kirkwood integration method. *The Journal of Chemical Physics*, 117(10):4947, 2002. Later Vega paper comments that this is a seminal paper demonstrating it is possible to compute solubilities from computer models.

[45] A. Fredenslund, R. L. Jones, and J. M. Prausnitz. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE J.*, 21(6):1086–1099, Nov. 1975.

[46] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, and D. T. Mainz. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for ProteinLigand Complexes. *J Med Chem*, 49(21):6177–6196, Oct. 2006.

[47] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery Jr, T. Vreven, K. N. Kuden, J. C. Burant, J. M. Milliam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomberts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. *Gaussian03*. Gaussian, Inc., Wallingford CT, c.02 edition, Sept. 2004.

[48] E. Gallicchio, N. Deng, P. He, A. L. Perryman, D. N. Santiago, S. Forli, A. J. Olson, and R. M. Levy. Virtual screening of integrase inhibitors by large scale binding free energy calculations. *J Comput Aided Mol Des*, 2014.

[49] E. Gallicchio, M. Lapelosa, and R. M. Levy. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of ProteinLigand Binding Affinities. *Journal of Chemical Theory and Computation*, 6(9):2961–2977, Sept. 2010.

[50] E. Gallicchio, M. Lapelosa, and R. M. Levy. The Binding Energy Distribution Analysis Method (BEDAM) for the Estimation of Protein-Ligand Binding Affinities. *J Chem Theory Comput.*, 6(9):2961–2977, Sept. 2010.

[51] M. Geballe. Overview of SAMPL Hydration Challenge. In *SAMPL 2011*, Stanford, CA, Aug. 2011.

[52] M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, and P. J. Taylor. The SAMPL2 blind prediction challenge: introduction and overview. *Journal of Computer-Aided Molecular Design*, 24(4):259–279, Apr. 2010.

[53] A. M. Geretti, D. Armenia, and F. Ceccherini-Silberstein. Emerging patterns and implications of HIV-1 integrase inhibitor resistance. *Current Opinion in Infectious Diseases*, 25(6):677–686 10.1097–QCO.0b013e32835a1de7, 2012.

[54] M. K. Gilson and H.-X. Zhou. Calculation of Protein-Ligand Binding Affinities. *Ann. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.

[55] J. Gmehling, J. Li, and M. Schiller. A modified unifac model. 2. present parameter matrix and results for di fferent thermodynamic properties. *Industrial & Engineering Chemistry Research*, 32(1):178–193, 1993.

[56] J. Gmehling, J. Lohmann, A. Jakob, J. Li, and R. Joh. A Modified UNIFAC (Dortmund) Model. 3. Revision and Extension. *Ind. Eng. Chem. Res.*, 37(12):4876–4882, Dec. 1998.

[57] J. Greenwald, V. Le, S. L. Butler, F. D. Bushman, and S. Choe. The Mobility of an HIV-1 Integrase Active Site Loop Is Correlated with Catalytic Activity,. *Biochemistry*, 38(28):8892–8898, 1999.

[58] J. P. Guthrie. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *The Journal of Physical Chemistry B*, 113(14):4501–4507, 2009.

[59] J. P. Guthrie. Overview of Experimental Data for Hydration. In *SAMPL 2011*, Stanford, CA, Aug. 2011.

[60] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, Aug 2008.

[61] H. K. Hansen, P. Rasmussen, A. Fredenslund, M. Schiller, and J. Gmehling. Vapor-liquid equilibria by UNIFAC group contribution. 5. Revision and extension. *Ind. Eng. Chem. Res.*, 30(10):2352–2355, Oct. 1991.

[62] S. Hare, G. N. Maertens, and P. Cherepanov. 3[prime]-Processing and strand transfer catalysed by retroviral integrase in crystallo. *EMBO J*, 31(13):3020–3028, 2012.

[63] S. Hare, S. J. Smith, M. Métifiot, A. Jaxa-Chamiec, Y. Pommier, S. H. Hughes, and P. Cherepanov. Structural and Functional Analyses of the Second-Generation Integrase Strand Transfer Inhibitor Dolutegravir (S/GSK1349572). *Molecular Pharmacology*, 80(4):565–572, Oct. 2011.

[64] S. Hare, A. M. Vos, R. F. Clayton, J. W. Thuring, M. D. Cummings, and P. Cherepanov. Molecular mechanisms of retroviral integrase inhibition and the evolution of viral resistance. *Proceedings of the National Academy of Sciences*, 107(46):20057–20062, Nov. 2010.

[65] P. C. D. Hawkins and A. Nicholls. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model*, 52(11):2919–2936, Nov. 2012.

[66] P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, and M. T. Stahl. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model*, 50(4):572–584, Apr. 2010.

[67] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations . *Journal of Computational Chemistry*, 18(2):1463–1472, Sept. 1997.

[68] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[69] B. Hess and N. F. A. van der Vegt. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *The Journal of Physical Chemistry B*, 110(35):17616–17626, Sept. 2006.

[70] M. Hewitt, M. T. D. Cronin, S. J. Enoch, J. C. Madden, D. W. Roberts, and J. C. Dearden. In SilicoPrediction of Aqueous Solubility: The Solubility Challenge. *J. Chem. Inf. Model.*, 49(11):2572–2587, Nov. 2009.

[71] H. Hogues, T. Sulea, and E. O. Purisima. Exuastive docking and solvated interaction energy scoring: Lessons learned from the SAMPL4 challenge. *J Comput Aided Mol Des*, 2014.

[72] N. Homeyer and H. Gohlke. FEW-A workflow tool for free energy calculations of ligand binding. *J Comput Chem*, 34(11):965–973, Jan. 2013.

[73] A. J. Hopfinger, E. X. Esposito, A. Llinàs, R. C. Glen, and J. M. Goodman. Findings of the Challenge To Predict Aqueous Solubility. *J. Chem. Inf. Model.*, 49(1):1–5, Jan. 2009.

[74] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *The Journal of Chemical Physics*, 120(20):9665, 2004.

[75] L. D. Hughes, D. S. Palmer, F. Nigsch, and J. B. O. Mitchell. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *J. Chem. Inf. Model.*, 48(1):220–232, Jan. 2008.

[76] P. Hünenberger and J. A. McCammon. Ewald artifacts in computer simulations of ionic solvation and ion–ion interaction: a continuum electrostatics study. *The Journal of Chemical Physics*, 110:1856, 1999.

[77] P. Hünenberger and M. Reif. *Single-Ion Solvation: Experimental and Theoretical Approaches to Elusive Thermodynamic Quantities*, volume 0 of *RSC Theoretical and Computational Chemistry Series2041-319X10.1039/2041-319Xhttp://ebook.rsc.org/?DOI=10.1039/2041-319X*. Royal Society of Chemistry, Cambridge, 2011.

[78] A. Jakalian, B. Bush, D. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.

[79] A. Jakalian, D. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, 2002.

[80] D. Japrung, U. Leartsakulpanich, S. Chusacultanachai, and Y. Yuthavong. Conflicting Requirements of Plasmodium falciparum Dihydrofolate Reductase Mutations Conferring Resistance to Pyrimethamine-WR99210 Combination. *Antimicrobial Agents and Chemotherapy*, 51(12):4356–4360, Dec. 2007.

[81] C. Jarzynski. Rare events and the convergence of exponentially averaged work values. *Physical Review E*, 73(4):46105, 2006.

[82] W. L. Jorgensen and J. Chandrasekhar. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79:926–935, 1983.

[83] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926, 1983.

[84] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, 118:11225–11236, 1996.

[85] K. A. Jurado, H. Wang, A. Slaughter, L. Feng, J. J. Kessl, Y. Koh, W. Wang, A. Ballandras-Colas, P. A. Patel, J. R. Fuchs, M. Kvaratskhelia, and A. Engelman. Allosteric integrase inhibitor potency is determined through the inhibition of HIV-1 particle maturation. *Proceedings of the National Academy of Sciences*, 110(21):8690–8695, May 2013.

[86] P. G. Karamertzanis, P. Raiteri, and A. Galindo. The Use of Anisotropic Potentials in Modeling Water and Free Energies of Hydration. *Journal of Chemical Theory and Computation*, 6(5):1590–1607, May 2010.

[87] M. Kastenholz and P. Hünenberger. Computation of methodology-independent ionic solvation free energies from molecular simulations. I. The electrostatic potential in molecular liquids. *The Journal of Chemical Physics*, 124:124106, 2006.

[88] M. Kastenholz and P. Hünenberger. Computation of methodology-independent ionic solvation free energies from molecular simulations. II. The hydration free energy of the sodium cation. *The Journal of Chemical Physics*, 124, 2006.

[89] C. Kehoe. Personal communication. Discussion of computed hydration free energies via Skype and e-mail, July 2011.

[90] C. W. Kehoe, C. J. Fennell, and K. A. Dill. Testing the SEA-water model of solvation free energies in the SAMPL 3 community blind test. *Journal of Computer-Aided Molecular Design*, Oct. 2011.

[91] D. D. Kemp and M. S. Gordon. An Interpretation of the Enhancement of the Water Dipole Moment Due to the Presence of Other Water Molecules. *The Journal of Physical Chemistry A*, 112(22):4885–4894, June 2008.

[92] J. J. Kessl, N. Jena, Y. Koh, H. Taskent-Sezgin, A. Slaughter, L. Feng, S. de Silva, L. Wu, S. F. J. Le Grice, A. Engelman, J. R. Fuchs, and M. Kvaratskhelia. Multimode, Cooperative Mechanism of Action of Allosteric HIV-1 Integrase Inhibitors. *Journal of Biological Chemistry*, 287(20):16801–16811, May 2012.

[93] P. Klimovich and D. L. Mobley. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *Journal of Computer-Aided Molecular Design*, 24(4):307–316, 2010.

[94] J. L. Knight and C. L. Brooks III. Surveying implicit solvent models for estimating small molecule absolute hydration free energies. *Journal of Computational Chemistry*, 32(13):2909–2923, July 2011.

[95] L. Krishnan and A. Engelman. Retroviral Integrase Proteins and HIV-1 DNA Integration. *Journal of Biological Chemistry*, 287(49):40858–40866, Nov. 2012.

[96] B. Kuhn and P. A. Kollman. Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J Med Chem*, 43(20):3786–3791, Oct. 2000.

[97] J. L. Kulp, S. N. Blumenthal, Q. Wang, R. L. Bryan, and F. Guarnieri. A fragment-based approach to the SAMPL3 Challenge. *J Comput Aided Mol Des*, 26(5):583–594, Jan. 2012.

[98] A. Kumar and K. Y. J. Zhang. Computational fragment-based screening using RosettaLigand: the SAMPL3 challenge. *J Comput Aided Mol Des*, 26(5):603–616, Jan. 2012.

[99] I. D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences*, 96(18):9997–10002, Aug. 1999.

[100] J. Li, T. Zhu, G. D. Hawkins, P. Winget, D. A. Liotard, and D. G. Truhlar. Extension of the platform of applicability of SM5.42R universal solvation model, June 1999.

[101] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, pages NA–NA, 2010.

[102] S. Liu, Y. Wu, T. Lin, R. Abel, J. P. Redmann, C. M. Summa, V. R. Jaber, N. M. Lim, and D. L. Mobley. Lead optimization mapper: automating free energy calculations for lead optimization. *J Comput Aided Mol Des*, 27(9):755–770, Sept. 2013.

[103] N. Lu, D. A. Kofke, and T. B. Woolf. Improving the efficiency and reliability of free energy perturbation calculations using overlap sampling methods. *J Comput Chem*, 25(1):28–40, 2004.

[104] J. Luccarelli, J. Michel, J. Tirado-Rives, and W. L. Jorgensen. Effects of Water Placement on Predictions of Binding Affinities for p38α MAP Kinase Inhibitors. *J Chem Theory Comput.*, 6(12):3850–3856, Jan. 2010.

[105] J. L. MacCallum and D. P. Tieleman. Calculation of the water-cyclohexane transfer free energies of neutral amino acid side-chain analogs using the OPLS all-atom force field. *Journal of Computational Chemistry*, 24(15):1930–1935, Sept. 2003.

[106] G. N. Maertens, S. Hare, and P. Cherepanov. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature*, 468(7321):326–329, 2010.

[107] C. Manly, J. Chandrasekhar, J. Ochterski, J. Hammer, and B. Warfield. Strategies and tactics for optimizing the Hit-to-Lead process and beyond—A computational chemistry perspective. *Drug Discovery Today*, 13(3-4):99–109, Feb. 2008.

[108] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Performance of sm6, sm8, and smd on the sampl1 test set for the prediction of small-molecule solvation free energies. *The Journal of Physical Chemistry B*, 113(14):4538–4543, 2009. PMID: 19253989.

[109] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B*, 113(18):6378–6396, May 2009.

[110] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comp. Chem.*, 30(13):2157–2164, Oct. 2009.

[111] MarvinSketch. 2013.

[112] H. Matter, E. Defossa, U. Heinelt, P.-M. Blohm, D. Schneider, A. Müller, S. Herok, H. Schreuder, A. Liesum, V. Brachvogel, P. Lönze, A. Walser, F. Al-Obeidi, and P. Wildgoose. Design and quantitative structureactivity relationship of 3-amidinobenzyl-1h-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor xa. *Journal of Medicinal Chemistry*, 45(13):2749–2769, 2002. PMID: 12061878.

[113] B. Mennucci, E. Cancès, and J. Tomasi. Evaluation of Solvent Effects in Isotropic and Anisotropic Dielectrics and in Ionic Solutions with a Unified Integral Equation Method: Theoretical Bases, Computational Implementation, and Numerical Applications. *The Journal of Physical Chemistry B*, 101(49):10506–10517, Dec. 1997.

[114] M. Métifiot, K. Maddali, B. C. Johnson, S. Hare, S. J. Smith, X. Z. Zhao, C. Marchand, T. R. Burke, S. H. Hughes, P. Cherepanov, and Y. Pommier. Activities, Crystal Structures, and Molecular Dynamics of Dihydro-1H-isoindole Derivatives, Inhibitors of HIV-1 Integrase. *ACS Chemical Biology*, 8(1):209–217, 2013.

[115] J. Michel and J. W. Essex. Hit identification and binding mode predictions by rigorous free energy simulations. *Journal of Medicinal Chemistry*, 51(21):6654–6664, 2008.

[116] J. Michel and J. W. Essex. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *Journal of Computer-Aided Molecular Design*, pages 1–20, 2010.

[117] J. Michel, N. Foloppe, and J. W. Essex. Rigorous Free Energy Calculations in Structure-Based Drug Design. *Mol Inf.*, 29(8-9):570–578, July 2010.

[118] J. Michel, J. Tirado-Rives, and W. Jorgensen. Prediction of the Water Content in Protein Binding Sites. *The Journal of Physical Chemistry B*, 2009.

[119] J. Michel, M. L. Verdonk, and J. W. Essex. ProteinLigand Complexes: Computation of the Relative Free Energy of Different Scaffolds and Binding Modes. *J. Chem. Theory Comput.*, 3(5):1645–1655, Sept. 2007.

[120] D. L. Mobley, C. I. Bayly, M. D. Cooper, K. A. Dill, and K. A. Dill. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. *The Journal of Physical Chemistry B*, 113:4533–4537, 2009.

[121] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *Journal of Chemical Theory and Computation*, 5(2):350–358, 2009.

[122] D. L. Mobley, É. Dumont, J. D. Chodera, and K. Dill. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. *The Journal of Physical Chemistry B*, 111(9):2242–2254, 2007.

[123] D. L. Mobley, É. Dumont, J. D. Chodera, and K. A. Dill. Comparison of Charge Models for Fixed-Charge Force Fields: Small Molecule Hydration Free Energies in Explicit Solvent. *The Journal of Physical Chemistry B*, 115(5):1329–1332, Feb. 2011.

[124] D. L. Mobley, A. P. Graves, J. D. Chodera, A. McReynolds, B. K. Shoichet, and K. A. Dill. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *Journal of Molecular Biology*, 371:1118–1134, 2007.

[125] D. L. Mobley and P. V. Klimovich. Perspective: Alchemical free energy calculations for drug discovery. *The Journal of Chemical Physics*, 137(23):230901, 2012.

[126] D. L. Mobley, S. Liu, D. S. Cerutti, W. C. Swope, and J. E. Rice. Alchemical prediction of hydration free energies for SAMPL. *J Comput Aided Mol Des*, 26(5):551–562, 2012.

[127] D. L. Mobley, K. L. Wymer, and N. M. Lim. Blind prediction of solvation free energies from the SAMPL4 challenge. *J Comput Aided Mol Des*, 2014.

[128] M. Naïm, S. Bhat, K. N. Rankin, S. Dennis, S. F. Chowdhury, I. Siddiqi, P. Drabik, T. Sulea, C. I. Bayly, and A. Jakalian. Solvated interaction energy (SIE) for scoring protein-ligand binding affinities. 1. Exploring the parameter space. *J. Chem. Inf. Model*, 47(1):122–133, 2007.

[129] J. Newman, O. Dolezal, V. Fazio, T. Caradoc-Davies, and T. S. Peat. The DINGO dataset: a comprehensive set of data for the SAMPL challenge. *J Comput Aided Mol Des*, 26(5):497–503, May 2012.

[130] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, and V. S. Pande. Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *Journal of Medicinal Chemistry*, 51(4):769–779, Feb. 2008.

[131] OpenEye Python Toolkits. `http://www.eyesopen.com`, 2013.

[132] OpenEye Scientific Software, Inc. Santa Fe, NM, USA. Openeye unified python toolkit, 2012. version 2.0.0.

[133] P. N. Palma, M. J. Bonifácio, A. I. Loureiro, and P. Soares-da Silva. Computation of the binding affinities of catechol-O-methyltransferase inhibitors: Multisubstate relative free energy calculations. *J Comput Chem*, 33(9):970–986, Jan. 2012.

[134] D. S. Palmer, J. L. McDonagh, J. B. O. Mitchell, T. van Mourik, and M. V. Fedorov. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J Chem Theory Comput.*, page 120809113933003, Aug. 2012.

[135] K. Paton. An algorithm for finding a fundamental set of cycles of a graph. *Commun. ACM*, 12(9):514–518, Sept. 1969.

[136] D. A. Pearlman. A Comparison of Alternative Approaches to Free Energy Calculations. *J. Phys. Chem.*, 98(5):1487–1493, 1994.

[137] T. S. Peat, O. Dolezal, J. Newman, D. L. Mobley, and J. J. Deadman. Interrogating HIV integrase for compounds that bind – a SAMPL4 challenge. *J Comput Aided Mol Des*, 2014.

[138] T. S. Peat and G. Warren. Personal Communication. E-mail exchange, 2013.

[139] E. Perola, W. P. Walters, and P. S. Charifson. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, 56(2):235–249, Apr. 2004.

[140] A. L. Perryman, S. Forli, G. M. Morris, C. Burt, Y. Cheng, M. J. Palmer, K. Whitby, J. A. McCammon, C. Phillips, and A. J. Olson. A Dynamic Model of HIV Integrase Inhibition and Drug Resistance. *J. Mol. Biol.*, 397(2):600–615, 2010.

[141] P. K. Quashie, T. Mesplède, Y.-S. Han, T. Veres, N. Osman, S. Hassounah, R. Sloan, H.-T. Xu, and M. A. Wainberg. Biochemical analysis of the role of G118R-linked dolutegravir drug resistance substitutions in HIV-1 integrase. *Antimicrobial Agents and Chemotherapy*, Sept. 2013.

[142] P. K. Quashie, T. Mesplède, and M. A. Wainberg. Evolution of HIV integrase resistance mutations. *Current Opinion in Infectious Diseases*, 26(1):43–49 10.1097–QCO.0b013e32835ba81c, 2013.

[143] Y. Ran, N. Jain, and S. H. Yalkowsky. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *Journal of chemical information and computer sciences*, 41(5):1208–1217, Sept. 2001.

[144] C. H. Reynolds. Computer-aided drug design: a practical guide to protein-structure-based modeling. In K. M. Merz Jr, D. Ringe, and C. H. Reynolds, editors, *Drug Design: Structure- and Ligand-based Approaches*, pages 181–196. Cambridge University Press, 2010.

[145] R. F. Ribeiro, A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models. *J. Comput.-Aided Mol. Des.*, 24(4):317–333, Apr. 2010.

[146] G. J. Rocklin, D. L. Mobley, and K. A. Dill. Separated topologies—A method for relative binding free energy calculations using orientational restraints. *The Journal of Chemical Physics*, 138(8):085104, 2013.

[147] V. Schnecke and J. Bostrom. Computational chemistry-driven decision making in lead generation. *Drug Disc. Today*, 11(12):43–50, Feb. 2006.

[148] Schrödinger, LLC, New York, NY. Desmond; 3.4ed., 2012.

[149] Schrödinger, LLC, New York, NY. Canvas, 2013. version 1.6.

[150] Y. Shi, S. P. OConnor, D. Sitkoff, J. Zhang, M. Shi, S. N. Bisaha, Y. Wang, C. Li, Z. Ruan, R. M. Lawrence, H. E. Klei, K. Kish, E. C.-K. Liu, S. M. Seiler, L. Schweizer, T. E. Steinbacher, W. A. Schumacher, J. A. Robl, J. E. Macor, K. S. Atwal, and P. D. Stein. Arylsulfonamidopiperidone derivatives as a novel class of factor xa inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 21(24):7516 – 7521, 2011.

[151] M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129:124105, 2008. http://dx.doi.org/10.1063/1.2978177 See code at: https://simtk.org/home/pymbar.

[152] M. R. Shirts, D. Mobley, and S. P. Brown. Free-energy calculations in structure-based drug design. *Drug Design: Structure- and Ligand-Based Approaches*, 2010.

[153] M. R. Shirts and D. L. Mobley. An Introduction to Best Practices in Free Energy Calculations. In *Biomolecular Simulations*. Methods in Molecular Biology, 2013.

[154] M. R. Shirts, D. L. Mobley, J. D. Chodera, and V. S. Pande. Accurate and efficient corrections for missing dispersion interactions in molecular simulations. *The Journal of Physical Chemistry B*, 111(45):13052–13063, Nov. 2007.

[155] M. R. Shirts and V. S. Pande. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *The Journal of Chemical Physics*, 122:134508, 2005.

[156] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *The Journal of Chemical Physics*, 119(11):5740, 2003.

[157] D. Shivakumar, Y. Deng, and B. Roux. Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. *Journal of Chemical Theory and Computation*, 5(4):919–930, 2009.

[158] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley, and W. Sherman. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *Journal of Chemical Theory and Computation*, 6(5):1509–1519, May 2010.

[159] A. G. Skillman, G. L. Warren, and A. Nicholls. SAMPL at first glance: So much data, so little time..., 2008.

[160] A. W. Sousa da Silva and W. F. Vranken. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes*, 5(1):367–374, 2012.

[161] T. Steinbrecher. Free Energy Calculations in Drug Lead Optimization. In H. Gohlke, editor, *Protein-Ligand Interactions*. Wiley-VCH, 2012.

[162] T. Steinbrecher and A. Labahn. Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies. *Current medicinal chemistry*, 17(8):767–785, 2010.

[163] T. Steinbrecher, D. L. Mobley, and D. A. Case. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *The Journal of Chemical Physics*, 127(21):214108, Dec. 2007.

[164] T. Sulea, Q. Cui, and E. O. Purisima. Solvated Interaction Energy (SIE) for Scoring Protein–Ligand Binding Affinities. 2. Benchmark in the CSAR-2010 Scoring Exercise. *J. Chem. Inf. Model*, 51(9):2066–2081, Sept. 2011.

[165] T. Sulea, H. Hogues, and E. O. Purisima. Exhaustive search and solvated interaction energy (SIE) for virtual screening and affinity prediction. *J Comput Aided Mol Des*, 26(5):617–633, 2012.

[166] G. Surpateanu and B. I. Iorga. Evaluation of docking performance in a blinded virtual screening of fragment-like trypsin inhibitors. *J Comput Aided Mol Des*, 26(5):595–601, 2012.

[167] O. Trott and A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461, Jan. 2010.

[168] J. Truchon and C. I. Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling*, 47(2):488–508, 2007.

[169] M. Tsiang, G. S. Jones, A. Niedziela-Majka, E. Kan, E. B. Lansdon, W. Huang, M. Hung, D. Samuel, N. Novikov, Y. Xu, M. Mitchell, H. Guo, K. Babaoglu, X. Liu, R. Geleziunas, and R. Sakowicz. New Class of HIV-1 Integrase (IN) Inhibitors with a Dual Mode of Action. *Journal of Biological Chemistry*, 287(25):21189–21203, June 2012.

[170] M. van den Bosch, M. Swart, J. Snijders, H. Berendsen, A. E. Mark, C. Oostenbrink, W. van Gunsteren, and G. Canters. Calculation of the redox potential of the protein azurin and some mutants. *ChemBioChem*, 6(4):738–746, 2005.

[171] R. Varela, W. P. Walters, B. B. Goldman, and A. N. Jain. Iterative Refinement of a Binding Pocket Model: Active Computational Steering of Lead Optimization. *J Med Chem*, page 121009103244001, Oct. 2012.

[172] A. Villa and A. E. Mark. Calculation of the free energy of solvation for neutral analogs of amino acid side chains. *Journal of Computational Chemistry*, 23(5):548–553, 2002.

[173] A. Villa, R. Zangi, G. Pieffet, and A. E. Mark. Sampling and convergence in free energy calculations of protein-ligand interactions: the binding of triphenoxypyridine derivatives to factor Xa and trypsin. *Journal of Computer-Aided Molecular Design*, pages 673–686, 2003.

[174] A. R. D. Voet, A. Kumar, F. Berenger, and K. Y. J. Zhang. Combining in cerebra and in silico approaches for virtual screening and pose prediction in SAMPL4. *J Comput Aided Mol Des*, 2014.

[175] J. R. Votano, M. Parham, L. H. Hall, L. B. Kier, S. Oloff, A. Tropsha, Q. Xie, and W. Tong. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis*, 19(5):365–377, Sept. 2004.

[176] M. A. Wainberg, T. Mesplède, and P. K. Quashie. The development of novel HIV integrase inhibitors and the problem of drug resistance. *Current Opinion in Virology*, 2(5):656–662, 2012.

[177] J. Wang and H. Tingjun. Application of Molecular Dynamics Simulations in Molecular Property Prediction I: Density and Heat of Vaporization. *Journal of Chemical Theory and Computation*, 7(7):2151–2165, July 2011.

[178] J. Wang, W. Wang, P. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25:247–260, 2006.

[179] J. Wang, R. Wolf, J. Caldwell, P. Kollman, and D. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[180] L. Wang, B. J. Berne, and R. A. Friesner. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. *Proceedings of the National Academy of Sciences*, 109(6):1937–1942, Feb. 2012.

[181] L. Wang, Y. Deng, J. L. Knight, Y. Wu, B. Kim, W. Sherman, J. C. Shelley, T. Lin, and R. Abel. Modeling local structural rearrangements using fep/rest: Application to relative binding affinity predictions of cdk2 inhibitors. *Journal of Chemical Theory and Computation*, 9(2):1282–1293, 2013.

[182] R. Wang, L. Liu, L. Lai, and Y. Tang. SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J Mol Model*, 4(12):379–394, 1998.

[183] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. Lalonde S, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head. A Critical Assessment of Docking Programs and Scoring Functions. *Journal of Medicinal Chemistry*, 49:5912–5931, 2006.

[184] R. Wittig, J. Lohmann, and J. Gmehling. VaporLiquid Equilibria by UNIFAC Group Contribution. 6. Revision and Extension. *Ind. Eng. Chem. Res.*, 42(1):183–188, Jan. 2003.

[185] K. Yoshikawa, S. Kobayashi, Y. Nakamoto, N. Haginoya, S. Komoriya, T. Yoshino, T. Nagata, A. Mochizuki, K. Watanabe, M. Suzuki, H. Kanno, and T. Ohta. Design, synthesis, and sar of cis-1,2-diaminocyclohexane derivatives as potent factor xa inhibitors. part ii: Exploration of 66 fused rings as alternative s1 moieties. *Bioorganic & Medicinal Chemistry*, 17(24):8221 – 8233, 2009.

[186] K. Yoshikawa, A. Yokomizo, H. Naito, N. Haginoya, S. Kobayashi, T. Yoshino, T. Nagata, A. Mochizuki, K. Osanai, K. Watanabe, H. Kanno, and T. Ohta. Design, synthesis, and sar of cis-1,2-diaminocyclohexane derivatives as potent factor xa inhibitors. part i: Exploration of 56 fused rings as alternative s1 moieties. *Bioorganic & Medicinal Chemistry*, 17(24):8206 – 8220, 2009.

[187] R. J. Young, C. Adams, M. Blows, D. Brown, C. L. Burns-Kurtis, C. Chan, L. Chaudry, M. A. Convery, D. E. Davies, A. M. Exall, G. Foster, J. D. Harling, E. Hortense, S. Irvine, W. R. Irving, S. Jackson, S. Kleanthous, A. J. Pateman, A. N. Patikis, T. J. Roethka, S. Senger, G. J. Stelman, J. R. Toomey, R. I. West, C. Whittaker, P. Zhou, and N. S. Watson. Structure and property based design of factor xa inhibitors: Pyrrolidin-2-ones with aminoindane and phenylpyrrolidine p4 motifs. *Bioorganic & Medicinal Chemistry Letters*, 21(6):1582 – 1587, 2011.

[188] M. Zacharias, T. P. Straatsma, and J. A. McCammon. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *The Journal of Chemical Physics*, 100(12):9025–9031, 1994.

[189] B. Zagrovic and W. van Gunsteren. Computational analysis of the mechanism and thermodynamics of inhibition of phosphodiesterase 5A by synthetic ligands. *Journal of Chemical Theory and Computation*, 3(1):301–311, 2007.

[190] J. Zhang, F. J. Adrian, W. Jahnke, S. W. Cowan-Jacob, A. G. Li, R. E. Iacob, T. Sim, J. Powers, C. Dierks, F. Sun, G.-R. Guo, Q. Ding, B. Okram, Y. Choi, A. Wojciechowski, X. Deng, G. Liu, G. Fendrich, A. Strauss, N. Vajpai, S. Grzesiek, T. Tuntland, Y. Liu, B. Bursulaya, M. Azam, P. W. Manley, J. R. Engen, G. Q. Daley, M. Warmuth, and N. S. Gray. Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors. *Nature*, 463(7280):501–506, 2010.