JCTC Journal of Chemical Theory and Computation

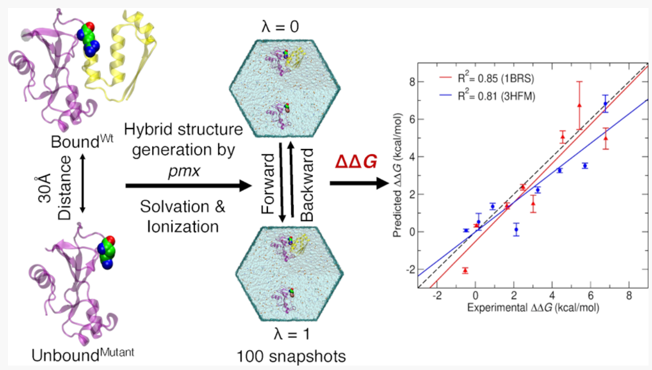
pubs.acs.org/JCTC Article

Implementing and Assessing an Alchemical Method for Calculating Protein–Protein Binding Free Energy

Dharmeshkumar Patel,* Jagdish Suresh Patel, and F. Marty Ytreberg*

Cite This: *J. Chem. Theory Comput.* 2021, 17, 2457–2464 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Protein–protein binding is fundamental to most biological processes. It is important to be able to use computation to accurately estimate the change in protein–protein binding free energy due to mutations in order to answer biological questions that would be experimentally challenging, laborious, or time-consuming. Although nonrigorous free-energy methods are faster, rigorous alchemical molecular dynamics-based methods are considerably more accurate and are becoming more feasible with the advancement of computer hardware and molecular simulation software. Even with sufficient computational resources, there are still major challenges to using alchemical free-energy methods for protein–protein complexes, such as generating hybrid structures and topologies, maintaining a neutral net charge of the system when there is a charge-changing mutation, and setting up the simulation. In the current study, we have used the *pmx* package to generate hybrid structures and topologies, and a double-system/single-box approach to maintain the net charge of the system. To test the approach, we predicted relative binding affinities for two protein–protein complexes using a nonequilibrium alchemical method based on the Crooks fluctuation theorem and compared the results with experimental values. The method correctly identified stabilizing from destabilizing mutations for a small protein–protein complex, and a larger, more challenging antibody complex. Strong correlations were obtained between predicted and experimental relative binding affinities for both protein–protein systems.

■ INTRODUCTION

Protein–protein binding is an essential phenomenon in molecular biology and directly mediates most functions in cells such as cellular metabolism, signal transduction, and coagulation among many other biological processes.[1,2] Mutations of the amino acids in protein–protein complexes can modulate or even disrupt protein–protein interactions by changing the associated binding free energy (ΔG) of the protein–protein complexes. The binding free energy of the protein–protein complexes determines the stability of association and the conditions for protein–protein complex formation.[3] It is important to be able to quantify the stabilities of protein complexes and how they can be modified by amino acid mutations and how they are affected by evolution.

Many techniques have been employed to determine the change in the protein–protein binding free energy due to a mutation (i.e., relative binding affinity, $\Delta\Delta G$). Experimental biophysical and biochemical methods are routinely used, but these methods are laborious, expensive, and time-consuming and are limited by technical challenges.[4−7] By contrast, computational methods can be relatively inexpensive, and the accuracy of such methods has been improved with the advancement of computational resources and better force fields.[8−10] Computational methods for estimating $\Delta\Delta G$ values can be broadly classified as either nonrigorous or rigorous.[11]

Nonrigorous free-energy methods typically use a single, static all-atom structure of the protein complex. These methods typically have energy functions that are trained using experimentally measured binding affinities or changes in affinities.[12,13] Many such semiempirical approaches have been developed that combine molecular mechanics and various optimized energy terms from available experimental data.[14] For example, BeAtMuSiC and mCSM use coarse-grained statistical potentials derived from known 3-D structures of proteins and machine learning.[15,16] FoldX uses empirical force field trained by experimentally measured binding free energies or changes in affinities.[12,13] The other so-called docking/scoring algorithms can predict binding affinities based on predicted binding poses

Received: October 4, 2020
Published: March 12, 2021

https://dx.doi.org/10.1021/acs.jctc.0c01045
J. Chem. Theory Comput. 2021, 17, 2457−2464

and putative binding interactions between protein−protein complexes.[17−19]

Rigorous free-energy approaches are based on the principles of statistical mechanics and use molecular simulations to explore the conformational space.[20] These methods typically provide more accurate $\Delta\Delta G$ predictions, compared to nonrigorous. One reason for this is that they inherently consider the conformational flexibility of the proteins and hence the entropic contribution. In recent years, rigorous approaches have made tremendous efficiency and theoretical advancements.[11,20] Rigorous free-energy calculation approaches are typically classified into three categories: endpoint methods, physical path sampling, and alchemical transformation.[20] Endpoint methods typically use molecular mechanics force fields with implicit solvent models such as molecular mechanics-generalized Born surface area (MMGB/SA) and molecular mechanics Poisson−Boltzmann surface area (MMPB/SA).[21,22] These methods are computationally less expensive than other rigorous approaches since simulations are only performed for two states; however, their accuracy is system-dependent and sensitive to simulation protocols such as sampling strategy and entropy calculation. For path sampling approaches, the physical unbinding and/or binding pathway of the protein with respect to its partner is sampled to obtain the underlying free-energy profile connecting bound and unbound states.[23−25] This category of methods can be very accurate but requires exhaustive conformational sampling along the pathway making it computationally expensive. Finally, alchemical methods exploit unphysical pathways by morphing, creating, and annihilating atoms.[26−29] These methods use molecular mechanics force fields as an energy function and the sampling of the correct thermodynamic ensemble is maintained by thermostatted and barostatted dynamics. The primary advantage is that the alchemical pathway does not need to be correlated with the physical binding process. This is particularly advantageous when considering relative binding affinity calculations due to single amino acid mutations (such as the current study). In this case, one needs to only calculate the free-energy change due to alchemically mutating the amino acid to another type in both the bound and unbound states.

Rigorous molecular dynamics (MD)-based alchemical free-energy calculation can be performed using equilibrium (e.g., free-energy perturbation,[30] thermodynamics integration[31]) or nonequilibrium (e.g., the Jarzynski equality,[32,33] Crooks fluctuation theorem[34]) methods. The initial simulation setup is the same for both equilibrium and nonequilibrium methods, but the protocols used during the simulations and postanalyses are different. The Hamiltonian H is coupled to a parameter λ that navigates the system from wild-type ($\lambda = 0$) to mutant ($\lambda = 1$). While such alchemical methods can be very accurate, they can also be computationally expensive since sufficient sampling is required to overcome the energetic and entropic barriers. In addition, the initial setup is not user-friendly, particularly when there is a change in the net charge of the system.[29,35,36] Specifically, the setup requires the topology of the protein system to ensure that all bonded and nonbonded interactions are correctly switched from $\lambda = 0$ to 1.

To enable more user-friendly alchemical free-energy calculations, de Groot et al. developed a package called *pmx* that automatically generates hybrid protein structures and topologies using force field-specific pregenerated mutation libraries.[37−39] Moreover, to maintain the net charge of the system during alchemical transformation, they developed an approach that uses two protein systems in a single simulation box (double-system/single-box).[37,40] Their approach of using *pmx*-generated topologies with a double-system/single-box approach was previously used to predict protein folding $\Delta\Delta G$ values due to mutations.[37,38] Prior to the development of the *pmx package*, de Groot et al. used the hybrid topology approach to calculate binding free energies for ubiquitin in complex with different protein substrates using a fast-growth thermodynamic integration approach with the Crooks−Gaussian intersection (CGI) method.[41] The main purpose of their study was to analyze ubiquitin conformations due to point mutations and predict the sign of $\Delta\Delta G$ for binding different substrates. They studied 11 mutations and obtained a Pearson correlation coefficient of 0.70 ($p = 0.016$). However, they have not explored the transition time per snapshot for nonequilibrium simulations. Later, the same group tested *pmx* with double-system/single-box approach to predict $\Delta\Delta G$ binding free energies for the protein−protein complex of α-chymotrypsin with its inhibitor Turkey Ovomucoid third domain with nine observed mutations of site L18 of Turkey Ovomucoid third domain.[40] The correlation coefficient between predicted and experimental $\Delta\Delta G$ was 0.80. Although promising, this protein−protein complex is small, all nine mutations occurred at the same amino acid site and were noncharge mutations.

Here, we tested the performance of using *pmx* with a double-system/single-box approach in a systematic manner using two protein−protein complexes of different sizes with a wide range of experimental $\Delta\Delta G$ values. For each system, we selected eight mutations from different sites with a broad range of experimental $\Delta\Delta G$ values. We estimated $\Delta\Delta G$ values using *pmx* hybrid topologies with a double-system/single-box approach and the nonequilibrium CGI method. Predicted $\Delta\Delta G$ values were compared with experimental values. In contrast to previous studies by de Groot et al., we optimized the transition times for the most stabilizing and the most destabilizing mutations of each protein−protein system. Higher correlation was found for smaller protein−protein complex as well as the larger, more complex, antigen−antibody system. Our results suggest that there is still room for improvement in rigorous binding free-energy methods to reduce computational cost, especially for large, complex protein−protein systems.

■ METHODS

Test System Selection. We selected two protein−protein complexes from the SKEMPI database[42] as test systems for this study. We chose the relatively small Barnase (110 aa)−Barstar (89 aa) complex (Protein Data Bank (PDB) ID: 1BRS)[43] and the larger, more challenging, antigen−antibody complex of lysozyme (129 aa)−HY/HEL-10 FAB (429 aa) (PDB ID: 3HFM).[44] 1BRS has total 30 mutations, and 3HFM has 67 mutations reported with their binding constants (K_d) in SKEMPI database. We wanted to shortlist eight mutations from each system based on $\Delta\Delta G$ values. In order to do that we first calculated ΔG values for wild-type and mutant using the reported K_d and reported temperature (T) with eq 1

$$\Delta G = -RT \ln K_d \tag{1}$$

The $\Delta\Delta G$ values were calculated by taking the difference between ΔG of the mutant and ΔG of wild-type. The average $\Delta\Delta G$ value was used when multiple $\Delta\Delta G$ values for a single mutation were in the database (Supporting Information Table S1). We chose these systems and mutations based on several criteria: (i) $\Delta\Delta G$ values should vary in sign—important since

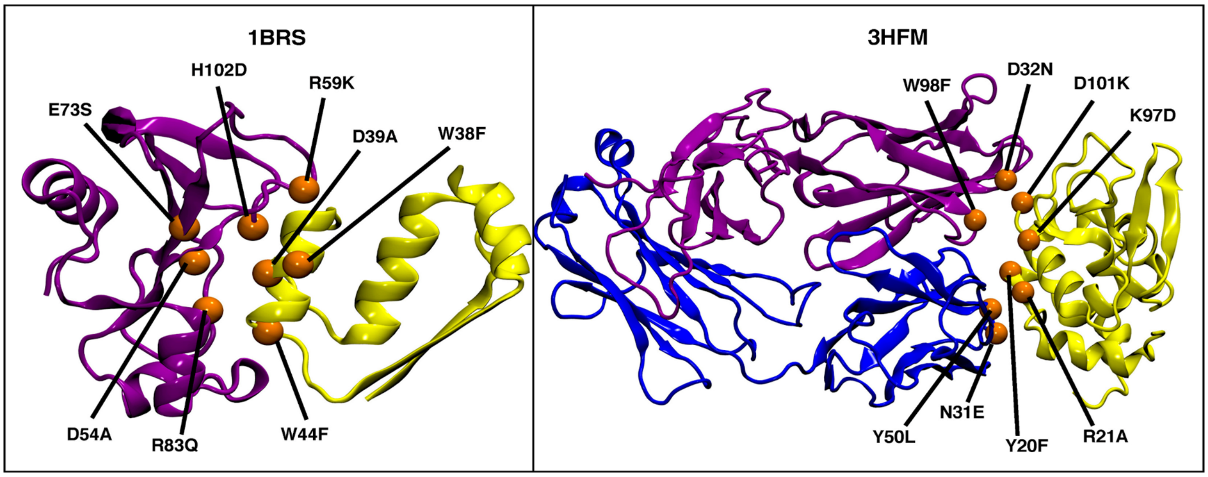


Figure 1. 3-D structures of the test systems used in the current study with the eight selected mutations shown as orange spheres. Left: Barnase (purple)−Barstar (yellow) protein complex (PDB ID: 1BRS); Right: lysozyme−HY (yellow) HEL-10 FAB (purple and blue) antigen−antibody complex (PDB ID: 3HFM).

mutations with negative (stabilizing) values are often more difficult to predict compared to positive (destabilizing) values; (ii) there should be a small number of missing residues in the 3-D structure of the protein complexes; (iii) chosen mutations should be nonalanine-scanning point mutations at differing amino acid sites; and (iv) reported mutations should be on multiple chains (Figure 1, Supporting Information Table S1).

Preparation of Protein−Protein Complexes. The 3-D structures of protein−protein complexes were downloaded from the PDB server (https://www.rcsb.org) and edited to preserve only the coordinates of the two or three interacting chains listed in the SKEMPI database.[42] All missing residues and atoms were then added using MODELLER software.[45] Mutants were generated using the BuildModel command from FoldX software.[12,13] This process provided nine input structures for each protein complex (a wild-type and eight mutant forms) to carry out alchemical free-energy calculations.

Construction of Hybrid Residues. Alchemical binding free-energy calculations require the construction of a non-physical pathway of intermediate states connecting the wild-type amino acid (λ = 0) to its mutant form (λ = 1). The *pmx* webserver[37,38] allows automatic generation of these intermediate states by producing hybrid amino acid states representing a mixture of wild-type and mutant form (see Figure 2). Both wild-

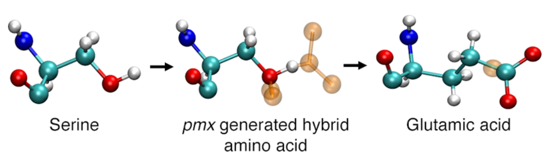


Figure 2. Example of a *pmx*-generated hybrid amino acid structure for serine (λ = 0) to glutamic acid (λ = 1). Dummy atoms are shown as transparent orange spheres.

type and mutant complex structure files were uploaded to the *pmx* webserver. The *pdb2gmx* option to add hydrogen atoms, and the Amber99SB*ILDN modified force field options were selected. The *pmx* webserver output consisted of hybrid structure and topology files compatible with GROMACS to perform the alchemical MD simulations.

Free-Energy Calculation and the Thermodynamic Cycle. To estimate relative binding free-energy values ($\Delta\Delta G$), we alchemically morphed the wild-type amino acids to their mutated forms (Figure 2). This process was replicated for both the bound and unbound states as indicated by horizontal arrows in the thermodynamic cycle shown in Figure 3. We can efficiently obtain ΔG_1 and ΔG_3 values with high

$$\begin{array}{ccc} \text{Bound}^{\text{Wt}} & \xrightarrow{\Delta G_1} & \text{Bound}^{\text{Mutant}} \\ \Delta G_4 \downarrow & \begin{aligned}\Delta\Delta G &= \Delta G_4 - \Delta G_2 \\ &= \Delta G_1 - \Delta G_3\end{aligned} & \downarrow \Delta G_2 \\ \text{Unbound}^{\text{Wt}} & \xrightarrow[\Delta G_3]{} & \text{Unbound}^{\text{Mutant}} \end{array}$$

Figure 3. Schematic representation of the thermodynamic cycle used to calculate relative binding free energies due to mutation ($\Delta\Delta G = \Delta G_1 - \Delta G_3$). Horizontal arrows indicate the non-physical pathways used in the current study where the amino acid was alchemically morphed from wild-type to its mutant form for both bound and unbound states.

accuracy using this approach.[46−48] By contrast, to carry out binding/unbinding simulations (vertical arrows in Figure 3), to calculate ΔG_2 and ΔG_4 values would be considerably more challenging and computationally expensive.

To estimate ΔG_1 and ΔG_3 (two horizontal arrows in Figure 3), we used the double-system/single-box approach developed by Gapsys et al.[40] Following this approach, we placed Bound$^{\text{Wt}}$ protein complex and Unbound$^{\text{Mutant}}$ protein in a single simulation box (λ = 0, Figure 4A) and similarly we placed Bound$^{\text{Mutant}}$ protein complex and Unbound$^{\text{Mutant}}$ protein in a second simulation box (λ = 1, Figure 4A). Figure 4B represents the series of steps involved for setting up the system for MD simulations and alchemical free-energy calculations. The distance between the two protein systems in each simulation box was maintained at 30 Å (Figure 4B) by applying position restraints on a single backbone atom close to the center of mass of each protein system. This separation distance was chosen to be larger than the short-range electrostatics cutoff to ensure that the two protein systems in a single simulation box did not interact with each other. Alchemical transformation from λ = 0 to 1 is termed "forward", where Bound$^{\text{Wt}}$ was transformed into Bound$^{\text{Mutant}}$ and simultaneously Unbound$^{\text{Mutant}}$ was transformed into Unbound$^{\text{Wt}}$, that is, "backward" λ = 1 to 0. Two

https://dx.doi.org/10.1021/acs.jctc.0c01045
J. Chem. Theory Comput. 2021, 17, 2457−2464

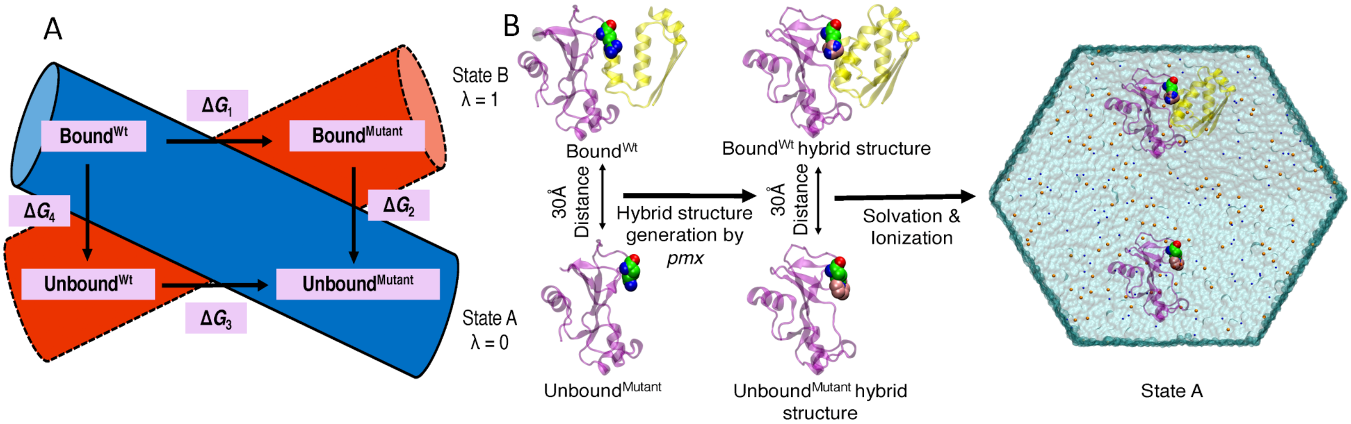


Figure 4. Double-system/single-box simulation setup. (A) Each colored cylinder represents a simulation box. During the forward alchemical transition, double systems consisting of Bound^{Wt} and Unbound^{Mutant} (blue cylinder, $\lambda = 0$) are morphed into Bound^{Mutant} and Unbound^{Wt} ($\lambda = 1$) states, respectively. Similarly, backward alchemical transition ($\lambda = 1$ to $\lambda = 0$) takes place in the red cylinder. (B) Schematic representation of the steps involved for setting up one of the double-system/single-box simulations for a mutation of 1BRS protein complex.

independent simulations (forward and backward) were thus performed to calculate the $\Delta\Delta G$ value for each mutation. Use of the double-system/single-box approach enabled us to maintain charge neutrality of the simulation system, even when an alchemical transformation involved a charge change between the wild-type and a mutant state, for example, R83Q.

MD Simulations and Alchemical Free-Energy Calculations. All MD simulations were carried out with the GROMACS-2018.3[49] MD simulation package using the Amber99SB*ILDN force field and the TIP3P water model.[50] The *pmx*-generated hybrid structures and modified force field files were used as an input. For each mutation, we prepared two simulation boxes ($\lambda = 0$ and $\lambda = 1$, Figure 4A) to carry out forward and backward transitions using the steps shown in Figure 4B. Both the states were solvated using dodecahedron water boxes. Na^+ and Cl^- ions were added at a 0.15 M concentration to neutralize the net charge. Both the simulation boxes were then energy-minimized for 10,000 steps using the steepest descent algorithm. Subsequent *NVT* followed by *NPT* ensemble simulations were performed for 500 ps for each simulation box. Note that in the scripts provided by *pmx*, *NVT* equilibration simulations were not performed; however, we included them in our study to reduce the system instability we observed. During the MD simulation, constant pressure and temperature were maintained using Parrinello−Rahman[51] pressure coupling at 1 atm and v-rescale temperature[52] coupling at 300 K. A 2 fs time step was used and each snapshot was saved at every 10 ps. Final production MD simulations were then performed for 40 ns to ensure sufficient sampling under *NPT* conditions. To prevent the diffusion of the proteins and maintain a 30 Å distance between the two protein systems, backbone carbons close to the center of mass were harmonically restrained with a force constant of 1000 kJ/mol nm^2. Choice of backbone C atoms used to apply position restraints for 1BRS was made based on the bound and unbound forms: (i) site A40 of bound-state Barstar; (ii) site A74 of unbound Barnase; and (iii) site L20 of unbound Barstar. While for 3HFM, (i) site Q37 of the bound-state light chain; (ii) site H41 of unbound state of the light chain; and (iii) site L56 of the antigen. The light chain is always bound to the heavy chain regardless of whether the antigen is bound or unbound. These positional restraints affect only the translational degrees of freedom of the proteins, not the overall structure or orientation of the proteins. The contribution of the positional restraints to the estimation of ΔG will be the same for the bound and unbound form of the proteins and thus the bias cancels out when calculating $\Delta\Delta G$, as is the case for the current study.

After the equilibrium MD simulations, fast-growth nonequilibrium alchemical simulations were performed to estimate the $\Delta\Delta G$. From each equilibrated MD simulation, the first 10 ns of the trajectory was discarded, and the last 30 ns was used to generate 100 snapshots (i.e., every 300 ps). Each snapshot was used to initialize a nonequilibrium simulation with a transition time of 5 ns for 1BRS and 8 ns for 3HFM (see the Supporting Information) where λ was continuously changed from 0 to 1 or from 1 to 0. The speed of λ value change was set 2×10^{-7}/fs for all forward and backward transitions. The derivatives of the Hamiltonian with respect to λ were recorded at every step and free energies were calculated from the work (W) distributions obtained from integration according to eq 2.

$$W = \int_{\lambda=0}^{\lambda=1} \frac{\delta H}{\delta \lambda} \mathrm{d}\lambda \qquad (2)$$

$\Delta\Delta G$ was estimated by calculating the intersection of the forward and backward work distributions according to the CGI method as described in Goette and Grubmüller.[53] The scripts used for analysis and calculations of $\Delta\Delta G$ were obtained from the *pmx* package.

RESULTS AND DISCUSSION

The purpose of our study is to test the accuracy of using *pmx* hybrid topologies and alchemical free-energy calculations with the double-system/single-box approach developed by Gapsys et al. to estimate relative binding affinities of protein−protein complexes. The *pmx* package allows for automated generation of the necessary hybrid topologies that are otherwise challenging to generate, and the double-system/single-box approach is a simple approach to maintain a neutral charge even when a mutation changes the protein charge. We tested this approach on two protein−protein systems of varying sizes (1BRS and 3HFM). For each system, we selected eight distinct mutations with experimental $\Delta\Delta G$ values reported in the literature using the criteria listed under the Methods section.

For alchemical nonequilibrium free-energy calculations using the fast growth method,[39,54,55] the transition time from $\lambda = 0$ to 1 or $\lambda = 1$ to 0 significantly influences the accuracy of $\Delta\Delta G$

Table 1. Predicted Relative Binding Free Energy of Each Mutation of 1BRS at Different Transition Times between 1 to 5 ns for 100 Independent Transitions[a]

	$\Delta\Delta G$ (kcal/mol)							
mutations (1BRS)	experimental	1 ns	2 ns	3 ns	4 ns	5 ns	6 ns	7 ns
D54A	−0.53	17.71	15.77	11.94	4.92	−2.07	−2.46	−1.89
W44F	0.06 ± 0.2	0.48	0.23	0.41	0.61	0.32		
W38F	1.64 ± 0.2	0.94	1.14	1.13	1.02	1.38		
R59K	2.49	7.91	10.16	7.03	3.84	2.37		
E73S	3.01 ± 0.2	−27.01	−19.31	−5.3	−1.30	1.49		
H102D	4.55	12.39	10.98	9.86	7.87	5.05		
R83Q	5.42 ± 0.2	−2.39	13.59	15.18	9.35	6.73		
D39A	6.79	8.93	10.45	9.45	7.60	4.97	5.42	4.65

[a]Estimated $\Delta\Delta G$ values of all eight mutations of the 1BRS system for 100 independent transitions. The predicted $\Delta\Delta G$ values were compared with the corresponding experimental data. $\Delta\Delta G$ values beyond 5 ns of transition time are for test mutations D54A and D39A as a part of the convergence study.

Table 2. Predicted Relative Binding Free Energy of Each Mutation of 3HFM at Different Transition Times between 1 to 8 ns for 100 Independent Transitions[a]

	$\Delta\Delta G$ (kcal/mol)										
mutations (3HFM)	experimental	1 ns	2 ns	3 ns	4 ns	5 ns	6 ns	7 ns	8 ns	9 ns	10 ns
Y20F	−0.48	−7.53	−6.34	−3.45	−2.95	−0.34	−1.02	−0.83	0.07	−0.69	−0.98
D32N	0.17 ± 0.3	−3.47	−5.67	−2.99	−1.32	−1.59	−1.98	−1.14	0.53		
R21A	0.90	5.82	7.54	4.56	3.98	1.23	2.34	1.74	1.35		
D101K	2.13	16.98	13.27	7.43	3.59	−0.46	−1.27	0.87	0.12		
W98F	3.25 ± 0.16	7.49	5.89	6.78	2.35	−0.16	−0.87	0.46	2.23		
Y50L	4.39 ± 0.12	9.65	6.37	2.36	1.33	0.10	1.37	2.89	3.26		
N31E	5.71 ± 0.13	13.28	8.73	9.67	4.78	−1.26	0.56	2.34	3.52		
K97D	6.77 ± 0.14	10.25	10.47	7.09	5.36	9.00	7.20	8.33	6.83	7.86	8.24

[a]Estimated $\Delta\Delta G$ values of all eight mutations of the 3HFM system for 100 independent transitions. The predicted $\Delta\Delta G$ values were compared with the corresponding experimental data. $\Delta\Delta G$ values beyond 8 ns of transition time are for test mutations Y20F and K97D as a part of the convergence study.

prediction. Short transition times lead the system far away from the equilibrium leading to a heavily biased estimate, while long transition times are less biased but more computationally costly, so the right balance is required.[39] To develop our simulation protocol, we initially chose two mutations from the 1BRS and 3HFM as test cases. These cases represent the most stabilizing (1BRS:D54A, $\Delta\Delta G$ = −0.53 kcal/mol; 3HFM:Y20F, $\Delta\Delta G$ = −0.48 kcal/mol) and destabilizing (1BRS:D39A, $\Delta\Delta G$ = 6.79 kcal/mol; 3HFM:K97D, $\Delta\Delta G$ = 6.77 kcal/mol) charge-changing mutations from the list of eight selected mutations (See Tables 1 & 2). To determine a reasonable transition time for our production simulations, we calculated $\Delta\Delta G$ values for both the test case mutations of 1BRS and 3HFM using 100 snapshots with a range of transition times from 1 to 7 ns for 1BRS and 1 to 10 ns for 3HFM. Supporting Information Figure 1 shows that transition times of 5 ns for 1BRS and 8 ns for 3HFM were sufficient to accurately estimate the free energies for these challenging mutations.

$\Delta\Delta G$ values of the remaining six mutations of 1BRS and 3HFM were estimated using the optimized simulation protocol and the transition time established through test case mutations. The predicted $\Delta\Delta G$ values were within ±2 kcal/mol of experimental $\Delta\Delta G$ values for optimized transition times for both protein−protein systems. In addition, experimental $\Delta\Delta G$ errors are within ±0.2 kcal/mol for both the test systems.

Figure 5 shows the correlation between the predicted and experimental $\Delta\Delta G$ values for all mutations from both the test systems. The calculated $\Delta\Delta G$ values correlate well with experimental data (R^2 = 0.85) for a smaller system of 1BRS and (R^2 = 0.81) for the larger, antigen−antibody complex 3HFM. The noncharge mutations from the 1BRS system such as W44F and W38F have the predicted $\Delta\Delta G$ values within the range of ±0.5 kcal/mol of experimental $\Delta\Delta G$ values. The convergence time for these mutations was within 1−2 ns transition time/snapshot. In the case of 3HFM, the noncharge mutations Y20F, W98F, and Y50L have higher accuracy, within range of ±1 kcal/mol of experimental $\Delta\Delta G$ values compared to other charge-changing mutations. Conversely, the charge-changing mutations are challenging to achieve convergence in free-energy calculations with short transition time. Longer

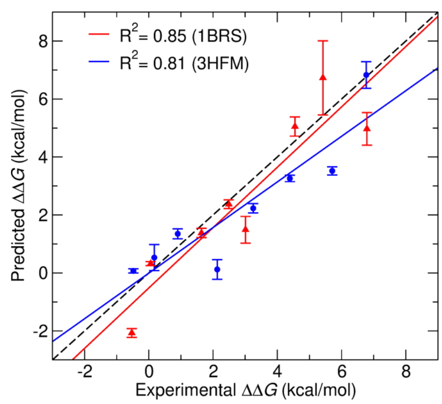


Figure 5. Correlation between predicted and experimental $\Delta\Delta G$ values for 1BRS (red) and 3HFM (blue) systems. The dashed black line shows perfect correlation.

https://dx.doi.org/10.1021/acs.jctc.0c01045
J. Chem. Theory Comput. 2021, 17, 2457−2464

transition times are likely needed in these cases to allow for sufficient conformational sampling. All the charge-changing mutations of the 1BRS system converged at around a 5 ns transition time with relatively high accuracy (±2 kcal/mol of experimental ΔΔ*G*). However, in 3HFM, the charge-changing mutations show convergence at around 8 ns transition time with an accuracy of ±2.5 kcal/mol of experimental ΔΔ*G*.

Both the test systems in this study were previously used by our laboratory to predict ΔΔ*G* values for the same eight mutations using the nonrigorous methods FoldX and MD + FoldX and rigorous coarse-grained umbrella sampling MD simulations.[56] The *pmx* with a double-system/single-box approach significantly outperforms the accuracy our previous FoldX[12,13] (1BRS:R^2 = 0.59, 3HFM:R^2 = −0.005), MD + FoldX[57−59] (1BRS:R^2 = 0.62, 3HFM:R^2 = 0.04), and coarse-grained umbrella sampling (1BRS:R^2 = 0.85, 3HFM:R^2 = 0.35) estimates in both the complexes. There is an especially large improvement in the accuracy of predicted ΔΔ*G* values for the antigen−antibody complex, 3HFM, with all-atom *pmx* with a double-system/single-box approach.

In this study, we used 100 snapshots per mutation to initiate the alchemical transitions and each snapshot was simulated for 5 ns. This means that 500 ns total simulation time was used to estimate ΔΔ*G* for both forward and backward directions. The equilibration simulation required ∼4500 CPUh for one mutation for the 1BRS system while in the case of 3HFM, it required ∼85,300 CPUh. With *pmx* with a double-system/single-box approach, the alchemical nonequilibrium simulation time is the major contributing factor to estimate the computational cost for the calculation of one ΔΔ*G*. In the 1BRS system, nonequilibrium simulations required ∼45,000 CPUh for 100 transitions per ΔΔ*G* prediction, however almost 30 times more CPUh (∼1,364,800) required in the case of the 3HFM system. It should also be noted that nonequilibrium alchemical transition is trivially parallelizable in that each of the 100 transitions can be run independently without relying on the completion of the previous simulation.

In order to obtain accurate binding free-energy values for protein−protein complex, exhaustive conformational sampling is required in order to sufficiently explore conformational space. Larger protein−protein complexes, such as antigen−antibody complex 3HFM studied here, require longer simulations to obtain convergence compared to smaller protein−protein complexes such as 1BRS.[60−62] In our study, we first optimized the protocol to calculate ΔΔ*G* values for the most stabilizing and the most destabilizing mutations of 1BRS and 3HFM systems and then applied the same protocol to rest of the mutations. We note that the accuracy of the nonequilibrium method could possibly be improved[39] via (i) longer equilibrium simulations to generate snapshots with more distant conformations, (ii) increasing the transition time per snapshot, and (iii) increasing number of independent transitions. We observed that in the case of 3HFM, the accuracy of ΔΔ*G* values was improved with increasing the transition time per snapshot.

Future work could involve using the alchemical double-system/single-box method but with coarse-grained protein models. Based on results from our previous study,[56] this may significantly reduce computational cost and still retain similar accuracy. However, coarse-grained hybrid topologies of the proteins have not yet been developed. Another approach to reducing computational cost could be use of a dual-resolution water model where water around the protein is atomistic and the rest of the water molecules coarse-grained.[63−65]

■ CONCLUSIONS

In this study, we have estimated protein−protein relative binding affinities due to single amino acid mutations using *pmx* hybrid topologies with a double-system/single-box approach. Nonequilibrium alchemical methods were used to generate ΔΔ*G* estimates for one small and one large protein−protein complex, and results were compared with experimental values. We obtained a significantly higher correlation between predicted and experimental ΔΔ*G* values for the small complex as well as the larger one. We were able to successfully distinguish stabilizing mutations from nonstabilizing mutations for all mutations in small complex and the large antigen−antibody complex. The accuracy of the predictions for the large complex is improved compared to previously tested rigorous and nonrigorous methods. Our results suggest that there are still potential areas for improvement in the reduction of computational cost for binding free-energy calculations, especially for larger protein−protein complexes. Future work could also be devoted to estimating binding free energies due to multiple mutations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c01045.

Experimental ΔΔG values of single mutations in the SKEMPI database and prediction of ΔΔ*G* values of test mutations of 1BRS and 3HFM systems as a function of transition time (PDF)

Scripts to set up and run the simulations for free-energy calculations (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Dharmeshkumar Patel − *Institute for Modeling Collaboration and Innovation, University of Idaho, Moscow, Idaho 83844, United States*; orcid.org/0000-0003-1601-256X; Email: dharmesh.patel5@gmail.com

F. Marty Ytreberg − *Institute for Modeling Collaboration and Innovation and Department of Physics, University of Idaho, Moscow, Idaho 83844, United States*; Email: ytreberg@uidaho.edu

Author

Jagdish Suresh Patel − *Institute for Modeling Collaboration and Innovation and Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844, United States*; orcid.org/0000-0003-4999-5347

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c01045

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported by the Complex for Modeling Complex Interactions sponsored by the NIGMS under award no. P20 GM104420 and by National Science Foundation EPSCoR Track-II grant under award number OIA1736253. Computer resources were provided in part by the Institute for Bioinformatics and Evolutionary Studies Computational Resources Core sponsored by the National Institutes of Health (grant no. P30 GM103324). This research also made use of the computational resources provided by the high-performance computing center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the U.S. Department of Energy (DOE) and the Nuclear Science User Facilities under contract no. DE-AC07-05ID14517. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

(1) Nooren, I. M. A.; Thornton, J. M. Diversity of Protein−Protein Interactions. *EMBO J.* **2003**, *22*, 3486−3492.

(2) Marsh, J. A.; Teichmann, S. A. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annu. Rev. Biochem.* **2015**, *84*, 551−575.

(3) Mosca, R.; Céol, A.; Aloy, P. Interactome3D: Adding Structural Details to Protein Networks. *Nat. Methods* **2013**, *10*, 47−53.

(4) Legrain, P.; Jestin, J.-L.; Schächter, V. From the Analysis of Protein Complexes to Proteome-Wide Linkage Maps. *Curr. Opin. Biotechnol.* **2000**, *11*, 402−407.

(5) Sprinzak, E.; Sattath, S.; Margalit, H. How Reliable are Experimental Protein-Protein Interaction Data? *J. Mol. Biol.* **2003**, *327*, 919−923.

(6) Mrowka, R.; Patzak, A.; Herzel, H. Is There a Bias in Proteome Research? *Genome Res.* **2001**, *11*, 1971−1973.

(7) von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S. G.; Fields, S.; Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **2002**, *417*, 399−403.

(8) D'Annessa, I.; Di Leva, F. S.; La Teana, A.; Novellino, E.; Limongelli, V.; Di Marino, D. Bioinformatics and Biosimulations as Toolbox for Peptides and Peptidomimetics Design: Where Are We? *Front. Mol. Biosci.* **2020**, *7*, 66.

(9) Gumbart, J. C.; Roux, B.; Chipot, C. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* **2013**, *9*, 794−802.

(10) Kilburg, D.; Gallicchio, E. Recent Advances in Computational Models for the Study of Protein-Peptide Interactions. In *Advances in Protein Chemistry and Structural Biology*; Christov, C. Z., Ed.; Insights into Enzyme Mechanisms and Functions from Experimental and Computational Methods; Academic Press, 2016; Vol. *105*, pp 27−57.

(11) Siebenmorgen, T.; Zacharias, M. Computational prediction of protein-protein binding affinities. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. e1448.

(12) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More than 1000 Mutations. *J. Mol. Biol.* **2002**, *320*, 369−387.

(13) Schymkowitz, J. W. H.; Rousseau, F.; Martins, I. C.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. Prediction of Water and Metal Binding Sites and Their Affinities by Using the Fold-X Force Field. *PNAS* **2005**, *102*, 10147−10152.

(14) Li, M.; Petukh, M.; Alexov, E.; Panchenko, A. R. Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. *J. Chem. Theory Comput.* **2014**, *10*, 1770−1780.

(15) Dehouck, Y.; Kwasigroch, J. M.; Rooman, M.; Gilis, D. BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res.* **2013**, *41*, W333−W339.

(16) Pires, D. E. V.; Ascher, D. B.; Blundell, T. L. MCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinformatics* **2014**, *30*, 335−342.

(17) Kastritis, P. L.; Bonvin, A. M. J. J. Are Scoring Functions in Protein−Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *J. Proteome Res.* **2010**, *9*, 2216−2225.

(18) Gromiha, M. M.; Yugandhar, K.; Jemimah, S. Protein-protein interactions: scoring schemes and binding affinity. *Curr. Opin. Struct. Biol.* **2017**, *44*, 31−38.

(19) Pons, C.; Grosdidier, S.; Solernou, A.; Pérez-Cano, L.; Fernández-Recio, J. Present and future challenges and limitations in protein-protein docking. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 95−108.

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

(21) Chen, F.; Liu, H.; Sun, H.; Pan, P.; Li, Y.; Li, D.; Hou, T. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein-protein binding free energies and re-rank binding poses generated by protein-protein docking. *Phys. Chem. Chem. Phys.* **2016**, *18*, 22129−22139.

(22) Rastelli, G.; Rio, A. D.; Degliesposti, G.; Sgobba, M. Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **2010**, *31*, 797−810.

(23) Fu, H.; Cai, W.; Hénin, J.; Roux, B.; Chipot, C. New Coarse Variables for the Accurate Determination of Standard Binding Free Energies. *J. Chem. Theory Comput.* **2017**, *13*, 5173−5178.

(24) Patel, D.; Mahdavi, S.; Kuyucak, S. Computational Study of Binding of μ-Conotoxin GIIIA to Bacterial Sodium Channels NaVAb and NaVRh. *Biochemistry* **2016**, *55*, 1929−1938.

(25) Patel, D.; Kuyucak, S.; Doupnik, C. A. Structural Determinants Mediating Tertiapin Block of Neuronal Kir3.2 Channels. *Biochemistry* **2020**, *59*, 836−850.

(26) Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? In *Annual Reports in Computational Chemistry*; Spellmeyer, D. C., Wheeler, R., Eds.; Elsevier, 2007; Vol. 3, pp 41−59.

(27) *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer Series in Chemical Physics; Springer-Verlag: Berlin, Heidelberg, 2007.

(28) Gao, J.; Kuczera, K.; Tidor, B.; Karplus, M. Hidden Thermodynamics of Mutant Proteins: A Molecular Dynamics Analysis. *Science* **1989**, *244*, 1069−1072.

(29) Rocklin, G. J.; Mobley, D. L.; Dill, K. A.; Hünenberger, P. H. Calculating the Binding Free Energies of Charged Species Based on Explicit-Solvent Simulations Employing Lattice-Sum Methods: An Accurate Correction Scheme for Electrostatic Finite-Size Effects. *J. Chem. phys.* **2013**, *139*, 11B606_1.

(30) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.

(31) Straatsma, T. P.; Berendsen, H. J. C. Free Energy of Ionic Hydration: Analysis of a Thermodynamic Integration Technique to Evaluate Free Energy Differences by Molecular Dynamics Simulations. *J. Chem. Phys.* **1988**, *89*, 5876−5886.

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

(33) Jarzynski, C. Equilibrium Free-Energy Differences from Nonequilibrium Measurements: A Master-Equation Approach. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1997**, *56*, 5018−5035.

(34) Crooks, G. E. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Stat. Phys.* **1998**, *90*, 1481−1487.

(35) Öhlknecht, C.; Lier, B.; Petrov, D.; Fuchs, J.; Oostenbrink, C. Correcting electrostatic artifacts due to net-charge changes in the calculation of ligand binding free energies. *J. Comput. Chem.* **2020**, *41*, 986−999.

(36) Chen, W.; Deng, Y.; Russell, E.; Wu, Y.; Abel, R.; Wang, L. Accurate Calculation of Relative Binding Free Energies between Ligands with Different Net Charges. *J. Chem. Theory Comput.* **2018**, *14*, 6346−6358.

(37) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. *J. Comput. Chem.* **2015**, *36*, 348−354.

(38) Gapsys, V.; de Groot, B. L. Pmx Webserver: A User Friendly Interface for Alchemistry. *J. Chem. Inf. Model.* **2017**, *57*, 109−114.

(39) Aldeghi, M.; de Groot, B. L.; Gapsys, V. Accurate Calculation of Free Energy Changes upon Amino Acid Mutation. In *Computational Methods in Protein Evolution*; Springer, 2019; pp 19−47.

(40) Gapsys, V.; Michielssens, S.; Peters, J. H.; de Groot, B. L.; Leonov, H. Calculation of Binding Free Energies. In *Molecular Modeling of Proteins*; Springer, 2015; pp 173−209.
(41) Michielssens, S.; Peters, J. H.; Ban, D.; Pratihar, S.; Seeliger, D.; Sharma, M.; Giller, K.; Sabo, T. M.; Becker, S.; Lee, D.; Griesinger, C.; de Groot, B. L. A Designed Conformational Shift To Control Protein Binding Specificity. *Angew. Chem.* **2014**, *126*, 10535−10539.
(42) Jankauskaite, J.; Jiménez-García, B.; Dapkūnas, J.; Fernández-Recio, J.; Moal, I. H. SKEMPI 2.0: An Updated Benchmark of Changes in Protein−Protein Binding Energy, Kinetics and Thermodynamics upon Mutation. *Bioinformatics* **2019**, *35*, 462−469.
(43) Buckle, A. M.; Schreiber, G.; Fersht, A. R. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-.ANG. resolution. *Biochemistry* **1994**, *33*, 8878−8889.
(44) Padlan, E. A.; Silverton, E. W.; Sheriff, S.; Cohen, G. H.; Smith-Gill, S. J.; Davies, D. R. Structure of an Antibody-Antigen Complex: Crystal Structure of the HyHEL-10 Fab-Lysozyme Complex. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5938−5942.
(45) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinf.* **2016**, *54*, 5.6.1−5.6.37.
(46) Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical Free Energy Calculations for Drug Discovery. *J. Chem. Phys.* **2012**, *137*, 230901.
(47) Aleksandrov, A.; Thompson, D.; Simonson, T. Alchemical free energy simulations for biological complexes: powerful but temperamental.. *J. Mol. Recognit.* **2010**, *23*, 117−127.
(48) Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234−2246.
(49) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(50) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. phys.* **1983**, *79*, 926−935.
(51) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(52) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(53) Goette, M.; Grubmüller, H. Accuracy and Convergence of Free Energy Differences Calculated from Nonequilibrium Switching Processes. *J. Comput. Chem.* **2009**, *30*, 447−456.
(54) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chem., Int. Ed.* **2016**, *55*, 7364−7368.
(55) Seeliger, D.; de Groot, B. L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, *98*, 2309−2316.
(56) Patel, J. S.; Ytreberg, F. M. Fast Calculation of Protein-Protein Binding Free Energies Using Umbrella Sampling with a Coarse-Grained Model. *J. Chem. Theory Comput.* **2018**, *14*, 991−997.
(57) Miller, C. R.; Johnson, E. L.; Burke, A. Z.; Martin, K. P.; Miura, T. A.; Wichman, H. A.; Brown, C. J.; Ytreberg, F. M. Initiating a Watch List for Ebola Virus Antibody Escape Mutations. *PeerJ* **2016**, *4*, No. e1674.
(58) Patel, J. S.; Quates, C. J.; Johnson, E. L.; Ytreberg, F. M. Expanding the Watch List for Potential Ebola Virus Antibody Escape Mutations. *PLoS One* **2019**, *14*, No. e0211093.
(59) Yang, J.; Naik, N.; Patel, J. S.; Wylie, C. S.; Gu, W.; Huang, J.; Ytreberg, F. M.; Naik, M. T.; Weinreich, D. M.; Rubenstein, B. M. Predicting the Viability of Beta-Lactamase: How Folding and Binding Free Energies Correlate with Beta-Lactamase Fitness. *PLoS One* **2020**, *15*, No. e0233509.
(60) Kamisetty, H.; Ramanathan, A.; Bailey-Kellogg, C.; Langmead, C. J. Accounting for Conformational Entropy in Predicting Binding Free Energies of Protein-Protein Interactions. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 444−462.
(61) Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Böckmann, R. A. Predicting Free Energy Changes Using Structural Ensembles. *Nat. Methods* **2009**, *6*, 3−4.
(62) Cappel, D.; Hall, M. L.; Lenselink, E. B.; Beuming, T.; Qi, J.; Bradner, J.; Sherman, W. Relative Binding Free Energy Calculations Applied to Protein Homology Models. *J. Chem. Inf. Model.* **2016**, *56*, 2388−2400.
(63) Darré, L.; Tek, A.; Baaden, M.; Pantano, S. Mixing Atomistic and Coarse Grain Solvation Models for MD Simulations: Let WT4 Handle the Bulk. *J. Chem. Theory Comput.* **2012**, *8*, 3880−3894.
(64) Gonzalez, H. C.; Darré, L.; Pantano, S. Transferable Mixing of Atomistic and Coarse-Grained Water Models. *J. Phys. Chem. B* **2013**, *117*, 14438−14448.
(65) Patel, J. S.; Ytreberg, F. M. Calculation of Protein-Protein Binding Free Energies Using Umbrella Sampling with Dual Resolution Water Models. *Biophys. J.* **2019**, *116*, 291a.