Published in final edited form as:
J Phys Chem B. 2018 May 31; 122(21): 5389–5399. doi:10.1021/acs.jpcb.7b11367.

Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity Upon Mutation

Kyle A. Barlow[†], **Shane O Conchúir**[‡,¶], **Samuel Thompson**[§], **Pooja Suresh**[§], **James E. Lucas**[||], **Markus Heinonen**[⊥,#], and **Tanja Kortemme**[†,‡,¶,§,||,@]

[†]Graduate Program in Bioinformatics, University of California San Francisco, San Francisco, California, United States of America [‡]California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, United States of America [¶]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, United States of America [§]Graduate Program in Biophysics, University of California San Francisco, San Francisco, California, United States of America [||]Graduate Program in Bioengineering, University of California San Francisco, San Francisco, California, United States of America [⊥]Aalto University, Department of Computer Science, Espoo, Finland [#]Helsinki Institute for Information Technology HIIT, Helsinki, Finland [@]Chan Zuckerberg Biohub, San Francisco, CA 94158

Abstract

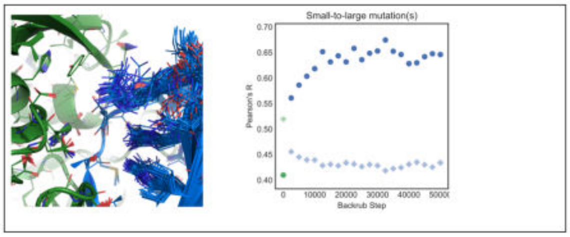
Computationally modeling changes in binding free energies upon mutation (interface $\Delta\Delta G$) allows large-scale prediction and perturbation of protein-protein interactions. Additionally, methods that consider and sample relevant conformational plasticity should be able to achieve higher prediction accuracy over methods that do not. To test this hypothesis, we developed a method within the Rosetta macromolecular modeling suite (flex ddG) that samples conformational diversity using "backrub" to generate an ensemble of models, then applying torsion minimization, side chain repacking and averaging across this ensemble to estimate interface $\Delta\Delta G$ values. We tested our method on a curated benchmark set of 1240 mutants, and found the method out-performed existing methods that sampled conformational space to a lesser degree. We observed considerable improvements with flex ddG over existing methods on the subset of small side chain to large side chain mutations, as well as for multiple simultaneous non-alanine mutations, stabilizing mutations, and mutations in antibody-antigen interfaces. Finally, we applied a generalized additive model (GAM) approach to the Rosetta energy function; the resulting non-linear reweighting model improved agreement with experimentally determined interface $\Delta\Delta G$ values, but also highlights the necessity of future energy function improvements.

Graphical abstract

Correspondence to: Tanja Kortemme.

Supporting Information Available
The following files are available as Supporting Information:

- suppinfo.pdf: Additional figures/tables
- flex-ddG-data.xlsx: Rosetta scores for all predictions



Introduction

Protein-protein interactions underlie essentially all biological processes, including signal transduction and antibody-antigen recognition. Many protein-protein interfaces are sensitive to mutations that can alter interaction affinity and specificity. In fact, mutations at protein-protein interfaces have been reported to be overrepresented within disease-causing mutations,[1] highlighting the central importance of these interactions to biology and human health. A sufficiently accurate computational method capable of predicting mutations that strengthen or weaken known protein-protein interactions would hence serve as a useful tool to dissect the role of specific protein-protein interactions in important biological processes. Coupled with state-of-the-art methods for protein engineering and design, such a method would also enhance our ability to create new and selective interactions, enabling the development of improved protein therapeutics, protein-based sensors, and protein materials.

Several prior methods have been developed to predict changes in protein-protein binding affinity upon mutation using different approaches to estimating energetic effects (scoring) and modeling structural changes (sampling). Common approaches include weighted energy functions that seek to describe physical interactions underlying protein-protein interactions, [2,3] statistical and contact potentials, [4–7] a combination of these approaches, [8,9] graph-based representations, [10] methods that sample backbone structure space locally around mutations, [11] and machine learning approaches. [12]

We set out to develop and assess methods for estimating experimentally determined changes in binding free energy after mutation (interface $\Delta\Delta G$) within the Rosetta macromolecular modeling suite. Rosetta is freely available for academic use, and allows combination of interface $\Delta\Delta G$ predictions with Rosetta's powerful protein design capabilities, which have proven successful in a variety of applications. [13,14] Prior projects have applied Rosetta predictions to dissect determinants of binding specificity and promiscuity, [15,16] enhance protein-protein binding affinities, [17,18] and to design modified [19] and new interactions, [20–22] but no prior benchmarking effort has quantitatively assessed the performance of predicting changes in binding free energy in Rosetta on a large, diverse benchmark dataset, in part because such datasets have only become available more recently. The current state-of-the-art Rosetta $\Delta\Delta G$ method, ddg_monomer,[23] has proven effective at predicting changes in stability of monomeric proteins after mutation, but had not yet been tested at predicting change of binding free energies in protein-protein complexes. Prior "computational alanine scanning" $\Delta\Delta G$ methods were benchmarked on mutations in protein-protein interfaces, focusing on mutations to alanine. [24–26] The original Rosetta alanine scanning method [24] did not sample backbone degrees of freedom, which is a first-order approximation for mutations to alanine (that are not expected to cause large backbone perturbations [27]), but less likely to be predictive for mutations to larger side chains which might require some degree of backbone rearrangement to accommodate the change. Inclusion of recent Rosetta energy function and sampling method developments, including methods that attempt to more aggressively sample conformational space, has not resulted in significant improvement to the alanine scanning method.[26]

We sought to create a method that would take into account aspects of the conformational plasticity of proteins by representing structures as an ensemble of individual full-atom models to explore biologically relevant and accessible portions of conformational space near the crystallographically determined input structures. Ensemble representations have previously been shown to be effective at predicting changes in protein stabilities after mutation [28] and at predicting the effects of mutation on protein-protein binding affinities, [29] as well as at improving $\Delta G_{binding}$ calculations between kinases and their inhibitors. [30]

We chose to sample conformational plasticity using the "backrub" protocol implemented in Rosetta.[31] The backrub method samples local side chain and backbone conformational changes, similar to those suggested to underlie observed conformational heterogeneity in high-resolution crystal structures, [32] and to accommodate evolved and designed mutations. [33] Backrub ensembles have been demonstrated to recapitulate properties of proteins that have been experimentally determined, such as side chain NMR order parameters, [34] tolerated sequence profiles at protein-protein [35] and protein-peptide interfaces, [36,37] and conformational variability between protein homologs. [38] Backrub has also proved effective in design applications, such as the redesign of protein-protein interfaces [19] and recapitulation of mutations that alter ligand-binding specificity. [39] When compared to ensembles generated via molecular dynamics simulations or the "PertMin" method, [40] backrub ensembles were shown to be the only ensembles capable of generating higher diversity (as measured by RMSD) between output models than from output models to the original input crystal structure. This observation suggests that backrub could be uniquely suited to produce diverse ensembles that effectively explore the local conformational space around an input structure. [40] Taken together, we hypothesized that these previously demonstrated properties of backrub ensembles would also make them an effective representation of near-native conformational states for use in predicting interface $\Delta\Delta G$ values.

Methods

Benchmark datasets

Developing and assessing the accuracy of a new method to predict changes in binding free energy after mutation requires a large and diverse benchmark set covering single mutations to all amino acid types, multiple mutations, and mutations across a variety of protein-protein interfaces. To facilitate comparisons to other methods and to avoid biases specific to our approach, we chose to use an existing benchmark dataset created by Dourado and Flores [11] during the development of their ZEMu (Zone Equilibration of Mutants) method. The ZEMu dataset was curated from the larger SKEMPI database [41] by avoiding a bias towards complexes in which a single position is repeatedly mutated, experimental data that are not peer-reviewed, redundancy (duplicate experimental values), mutations outside of interfaces, mutations involved in crystal contacts, and experimental $\Delta\Delta G$ values for which wild-type and mutant conditions (such as pH) varied. Confidence in the "known" experimental $\Delta\Delta G$ values is important, as it has been pointed out that the experimental methodology used can have a strong effect on the performance of predictors of changes in binding free energy. [42] The ZEMu dataset was also curated to include a range of both stabilizing and destabilizing mutants, small side chain to large side chain mutations, single and multiple mutations, and a diversity of complexes. Small-to-large mutations are defined as those dataset cases where all mutation(s) are at positions where the residue side chain increases in van der Waals volume post-mutation. [43]

After a review of the literature from which the known experimental $\Delta\Delta G$ values originated, we removed one data point from the 1254 point ZEMu set that we could not match to the originally reported affinity value. We also removed 5 mutations we determined to be duplicates, along with 8 mutations that were reverse mutations of other data points, leaving us with a test set of 1240 mutations (Table 1). We used SAbDab [44] to define complexes that contained at least one antibody binding partner. Our version of the ZEMu dataset is available in the Supporting Information as Dataset S1. All $\Delta\Delta G$ predictions described in the paper are available in the Supporting Information.

Rosetta implementation and prediction protocol

Our protocol, called "flex ddG", is implemented within the RosettaScripts interface to the Rosetta macro-molecular modeling software suite, [45] which makes the protocol easily adaptable to future improvements and energy function development. The method can be run using a Rosetta Scripts XML that is available in the Supporting Information as Listing 1. Version numbers of tested software are available in Table S1.

Flex ddG method steps are outlined in Fig. 1. **Step 1:** The protocol begins with an initial minimization (on backbone ϕ/ψ and side chain χ torsional degrees of freedom, using the limited-memory Broyden-Fletcher-Goldfarb-Shanno minimizer implementation within Rosetta, with Armijo inexact line search conditions (option "lbfgs_armijo_nonmonotone") of the input crystal structure of the wild-type protein complex. This (and later) minimizations are performed with harmonic restraints on pairwise atom distances to their values in the input crystal structure. Restraints were added for all pairs for C-α atoms within 9 Å of each other using a harmonic score potential defined to have the width (standard deviation) parameter set to 0.5 Å, and added to the Rosetta score function with a term weight of 1.0. Minimization is run until convergence (absolute score change upon minimization of less than one REU (Rosetta Energy Unit)). **Step 2:** Starting from the minimized input structure including both binding partners in the protein-protein complex, the backrub method in Rosetta[31] is used to create an ensemble of models. In brief, each backrub move is undertaken on a randomly chosen protein segment consisting of three to twelve adjacent residues in the neighborhood of any mutated position. The mutation neighborhood is defined by finding all residues in the protein-protein complex with a C-β atom (C-α for glycines) within 8 Å of any mutant position, then adding this residue and its adjacent N and C-terminal residues to the list of neighborhood residues. All atoms in the backrub segment are rotated locally about an axis defined as the vector between the endpoint C-α atoms. The allowed rotation angles for the backrub steps use Rosetta default values as described in Smith & Kortemme, 2008.[31] Backrub is run at a temperature of 1.2 kT, for up to 50,000 backrub Monte Carlo trials/steps (Table S2 shows that using a kT of 1.6 gives similar results to a kT of 1.2). Up to 50 output models are generated. **Step 3A:** For each of the 50 models in the ensemble output by backrub, the Rosetta "packer" is used to optimize side chain conformations for the wild-type sequence using discrete rotameric conformations [46] and simulated annealing. The packer is run with the multi-cool annealer option, [47] which is set to keep a history of the 6 best rotameric states visited during annealing. **Step 3B:** Independently and in parallel to step 3A, side chain conformations for the mutant sequence are optimized on all 50 models, introducing the mutation(s). **Step 4A:** Each of the 50 wild-type models is minimized, again adding pairwise interatomic distance restraints to the input structure. Minimization is run with the same parameters as in step 1; the coordinate restraints used in this step are taken from the coordinates of the Step 3A model. **Step 4B:** As Step 4A, but for each of the 50 mutant models. **Step 5A:** Each of the 50 minimized wild-type models are scored in complex, and the complex partners are scored individually. The scores of the split, unbound complex partners are obtained simply by moving the complex halves away from each other. No further minimization or side chain optimization is performed on the unbound partners before scoring. **Step 5B:** In the same fashion as Step 5A, each of the 50 minimized mutant models are scored in complex, and the complex partners are scored individually. **Step 6:** The interface $\Delta\Delta G$ score is calculated via Eq. 1 as the arithmetic mean over the different models produced:

$$\begin{aligned}\Delta\Delta G_{bind} &= \Delta G_{bind}^{MUT} - \Delta G_{bind}^{WT} \\ &= (\Delta G_{complex}^{MUT} - \Delta G_{partnerA}^{MUT} - \Delta G_{partnerB}^{MUT}) - (\Delta G_{complex}^{WT} - \Delta G_{partnerA}^{WT} - \Delta G_{partnerB}^{WT})\end{aligned} \quad (1)$$

We evaluate performance of the protocol by comparing predicted $\Delta\Delta G$ scores to known experimental values, using Pearson's correlation (R), Fraction Correct (FC), and Mean Absolute Error (MAE). Fraction Correct is defined as the number of cases in the dataset categorized correctly as stabilizing, neutral, or destabilizing, divided by the total number of cases in the dataset. Stabilizing mutations are defined as those with a $\Delta\Delta G <= -1.0$ kcal/mol, neutral as those with -1.0 kcal/mol $< \Delta\Delta G < 1.0$ kcal/mol, and destabilizing as those with $\Delta\Delta G >= 1.0$ kcal/mol.

MAE (Mean Absolute Error) is defined in Eq. 2 as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - x_i | = \frac{1}{n}\sum_{i=1}^{n} | e_i | \quad (2)$$

where y_i are the predicted $\Delta\Delta G$ values, x_i are the known, experimentally determined values, and e_i is the prediction error.

As a control, we ran the flex ddG protocol omitting the backrub ensemble generation step. This control protocol can in principle generate multiple models because of the minimization and packing steps, but in practice these models are structurally highly similar or identical.

Rosetta energy function

We utilized Rosetta's Talaris [46,48,49] all-atom energy function for the modeling steps. As we do not modify our models of the unbound state, several terms of the Rosetta energy function will cancel out in the final $\Delta\Delta G$ scoring because the ΔG of folding score of the unbound partners is subtracted from the total score of the complex (Eq. (1)). After subtraction, seven score terms remain, and combined, become the final interface $\Delta\Delta G$ score, dominated by solvation (fa_sol using an implicit solvation model [50]), hydrogen bonding and electrostatics [48,49,51] (hbond_sc: side chain-side chain hydrogen bonds; hbond_bb_sc: hydrogen bonds between backbone atoms and side chain atoms; hbond_lr_bb: long-range hydrogen bond interactions between backbone atoms; fa_elec: Coulomb electrostatics), and Lennard-Jones atomic packing interactions (fa_rep and fa_atr: repulsive and attractive components of the Lennard-Jones potential).

Score analysis

To investigate potential sources of prediction error on an individual score term basis, we used a generalized additive model [52] approach to fit Rosetta's predicted $\Delta\Delta G$ values to experimentally known values. First, we apply an unbiased logistic scaling to individual score terms,

$$h_{a,b}(x) = \frac{2e^{a}}{1 + e^{-xe^{b}}} - e^{a},$$

where a is the scaling range of the score, and b is the steepness of the sigmoid scaling. Both parameters are transformed through an exponential to ensure non-negativity. The scaling function h does not introduce bias, that is, $h_{\theta}(0) = 0$ for any θ. The scoring model results in a generalized additive model (GAM) over the M score terms,

$$f(\mathbf{x}) = \sum_{\mathbf{j=1}}^{\mathbf{M}} \mathbf{h}_{\mathbf{a_i},\,\mathbf{b_i}}(\mathbf{x})\,.$$

The parameters $\theta = (a_j, b_j)_{j=1}^{M}$ for the score terms were simultaneously sampled using a random walk Metropolis-Hastings MCMC algorithm (the `mhsample` function in Matlab) assuming a Gaussian likelihood as the target distribution

$$p(\theta; \mathbf{y}) = \mathbf{N}(\mathbf{f}(\mathbf{x_i}) \mid \mathbf{y_i}, \sigma_{\mathbf{n}}^{\mathbf{2}})$$

with a noise variance set to $\sigma_n^2 = 1.0$, and where $(\mathbf{x_i}, \mathbf{y_i})_{\mathbf{i=1}}^{\mathbf{N}}$ are the empirical observations y_i that correspond to the protein score terms $\mathbf{x_i}$, respectively. We sample for 1000 samples with a burn-in set to 1000 samples and a thinning parameter of 20. The proposal distribution was selected to be a symmetric uniform distribution such that $[a^{(s+1)}, b^{(s+1)}] \sim U(a^{(s)} \pm 2, b^{(s)} \pm 2)$. The resulting MCMC sample represents all logistics score scalings that reproduce the empirical measurements assuming an error model with noise variance σ_n^2.

Results and discussion

The overall performance of the protocol is summarized in Table 2. We compare 4 prediction methods: (a) our flex ddG backrub ensemble method, (b) the prior state-of-the-art Rosetta methodology, ddg_monomer,[23] (c) a control version of our flex ddG protocol which omits the backrub ensemble generation step, leaving only the minimization and packing steps, and (d) published data from the ZEMu (zone equilibration of mutants) method. [11] Data split by input protein-protein complex are shown in Table S3.

The new flex ddG method outperforms the comparison methods on the complete dataset in each of the correlation, MAE, and fraction correct metrics (Table 2). In particular, we see a large increase in performance relative to the other methods on the small-to-large subset of mutations. This is in accordance with our expectations that backrub ensembles should be able to sample small backbone conformational adjustments required to accommodate changes in amino acid residue size. Notably, application of backrub ensembles performs better than other methods that include backbone minimization steps only, including the current state-of-the-art Rosetta ddg_monomer method. On the small-to-large mutations subset, the ddg_monomer method achieves a Pearson correlation of only 0.31 compared to 0.65 with flex ddG.

Performance of the flex ddG method on the subset of single mutations to alanine is also competitive or outperforms the alternative methods. As we do not expect single mutations to alanine to require intensive backbone sampling, our method’s effectiveness on this subset shows that the method is fairly robust to the mutation type. As we chose to perform backrub sampling prior to introducing mutations, these results could suggest that flex ddG is effective by sampling underlying, relevant plasticity of the input crystal structure instead of distorting the local structure around a mutation to resolve a clash or poor interaction with a mutant side chain.

While the flex ddG method shows improved performance on the subset of multiple mutations as compared to the control and ddg_monomer methods, flex ddG did not match the performance of the ZEMu method on this subset. This result could indicate that further refinement of the backrub parameters is required when simultaneously sampling conformational space around the sites of multiple mutations. For example, while we modeled all mutations simultaneously, it is possible that a protocol that considers mutations sequentially could improve predictions. However, and remarkably, flex ddG outperforms ZEMu on the subset of cases with multiple mutations where none of the mutations are to alanine (Table 2). While any comments on the origins of this difference will be speculative especially with only limited structural information on the mutated proteins (as well as information on possible changed dynamics), we note that flex ddG predictions are more accurate for several cases in this dataset with experimental $\Delta\Delta G$ values around zero that ZEMu over-or underpredicts Finally, the flex ddG method also shows considerable improvements over other methods on the subset of antibody-antigen complexes (Table 2).

Fig. 2 illustrates the performance for the flex ddG and control methods on the complete dataset and small-to-large subsets using scatterplots comparing experimentally determined and computationally estimated changes in binding free energies for each of the cases in the datasets. In particular, a notable improvement with flex ddG over the control can be seen for the 13 small-to-large mutations that were experimentally determined to stabilize the protein-protein interface significantly ($\Delta\Delta G <= -1.0$ kcal/mol). For this set, the control method misclassifies most stabilizing mutations to have minimal effect or to be destabilizing (9 mutations with predicted Rosetta $\Delta\Delta G$ scores > 0) (Fig. 2d), whereas flex ddG identifies a sizable number (12 of 13 mutations) to have predicted Rosetta $\Delta\Delta G < 0$ (Fig. 2c), even though only one of these mutations is predicted to be strongly stabilizing (predicted $\Delta\Delta G$ score < -1). The capability to predict stabilizing mutations is especially important for challenging design applications to modulate binding affinity and selectivity, as well as creating entirely new high-affinity protein-protein interactions.

It has been previously observed that increasing the number of stabilizing mutations that are correctly identified (decreasing “false negatives”) might be accompanied by an increase in “false positives”, i.e. predictions that a mutation is stabilizing when it is not. However, using backbone ensembles was found to mitigate this effect by decreasing the number of false negatives more than it increases the number of false positives. [40] We therefore also evaluated the number of false positive predictions. For the complete dataset, there are 12 cases where the no backrub control method predicts a mutation incorrectly as stabilizing (Rosetta $\Delta\Delta G$ score $<= -1$) that were experimentally determined to destabilize the interface significantly ($\Delta\Delta G > 1$ kcal/mol). In contrast, flex ddG misclassified only 1 destabilizing mutation as stabilizing. We conclude that flex ddG makes both fewer false negative and fewer false positive predictions

In the following sections, we assess how different flex ddG implementations would affect prediction performance, focusing separately on sampling and scoring.

Effect of ensemble size

While the results presented above used an ensemble size of 50 members, we next investigated what the ideal ensemble size would be to maximize the predictive ability of our method. For example, prior methods used ensemble sizes ranging from ten [3] to thousands.[29] As the computational time required to run flex ddG increases linearly with ensemble size, determining an optimal size is practically relevant. We therefore evaluated the performance of flex ddG as we average across an increasing number of models (from 1 to 50, Fig. 3). The models are first sorted by the score of the corresponding repacked and minimized wild type model, such that producing a $\Delta\Delta G$ with 1 model will only use the lowest (best) scoring model, 2 models will use the 2 lowest scoring models, and so forth. Fig. 3(a) shows the performance on the complete dataset. As more models with increasing wild type complex score are averaged, correlation with known experimental values increases. Conversely, performance for the no backrub control method stays approximately constant as more models are averaged. This result indicates that sampling with backrub adds information that improves $\Delta\Delta G$ calculation even though the additional averaged models have higher scores (average ensemble total score is shown in Fig. S1). These higher scoring models would be excluded in methods such as the Rosetta ddg_monomer protocol, which typically use only the lowest scoring wild-type and mutant models. Similar observations on the utility of higher scoring models for stability prediction have been made previously. [53,54] Increasing the ensemble size may hence be useful to increase the odds of finding alternative conformations that are informative for estimating the effects of mutations, rather than simple minimization of structural models.

Instead of using just the three lowest energy models, [23] we find that the performance of the ddg_monomer method also improves as more output models are averaged (Fig. S2, Table S5). This was somewhat unexpected, as the no-backrub control method, which did not show an improvement with increasing ensemble size, is conceptually similar to the ddg_monomer method. However, the difference may arise from the fact that the ddg_monomer method ramps the weight of the repulsive Lennard-Jones term in the energy function during minimization. This strategy explores conformational space more broadly in different backbone ensemble members than minimization with a fully weighted repulsive term in the no-backrub control method. In this fashion, including more ensemble members generated by the ddg_monomer method increases the conformational plasticity sampled which in turn increases performance, as seen for the flex ddG method.

Using flex ddG, the subset of small-to-large mutations shows the largest increase in correlation with experimental $\Delta\Delta G$ values as more models are averaged (Fig. 3(b)). This result is consistent with our reasoning above that improved modeling of conformational plasticity is important for prediction performance, and that this effect is most important for significant changes in amino acid residue size. For the subset of multiple mutations where none are mutations to alanine (Fig. 3(c)), performance overall increases substantially initially when more models are added.

Averaging across increased numbers of models also improves correlation for the subset of single mutations to alanine (Fig. 3(d)). Here, improvements are seen up to averaging about 10 models, after which performance stays approximately constant. This observation indicates that increased sampling, in the very least, is not harmful for cases where one would expect structural changes to be relatively small on average.

To test whether the optimal number of models depends on the structural context of the mutation, we binned the complete dataset by secondary structure class (alpha-helix, strand, loop, turn) at the site of the mutation using DSSP. [55,56] For all secondary structure classes, we observed a performance increase when averaging over increasing number of models, reaching a plateau at around 20 to 30 models (Fig. S3). We observed a similar behavior when binning the dataset by residue burial at the site of mutation using solvent accessible surface area computed using DSSP[55,56] (Fig. S4). In all cases, we observed an increase in performance when averaging across a larger number of models.

In summary, from a practical standpoint, generating 20–30 models should constitute sufficient sampling for most cases. Sorting the generated models by score and selecting the best scoring 20–30 out of 50 models does not appear to be necessary, as not sorting the models by score (Fig. S5, Table S6) gives similar results to sorting the models (Fig. 3).

Effect of extent of backrub sampling in each trajectory

The extent of sampling can also be controlled by changing the number of Monte Carlo steps in the backrub simulations. Fig. 4 shows the effect of increasing the number of backrub Monte Carlo steps (while averaging all 50 models at each output step) on flex ddG performance, compared to a control method with zero backrub steps that uses only minimization and side chain packing. $\Delta\Delta G$ scores are calculated every 2,500 backrub steps.

After an initial increase for the first set of 2500 backrub steps, performance stays relatively constant for the complete dataset (Figure 4a) and for single mutations to alanine (Fig. 4d). However, for the subsets of small-to-large mutations (Figure 4b) and multiple mutations, none to alanine (Fig. 4c), performance increases considerably with increasing numbers of Monte Carlo steps. This increase in performance is similar to what was observed with averaging over more models for these subsets (Fig. 3b,c). Performance levels off at around 30,000 backrub Monte Carlo steps.

The increased performance does not appear to be simply a result of decreasing scores as the simulation progresses, as the average score of the minimized wild type complexes does not decrease uniformly across the sampled ensemble as the simulation progresses (Fig. S1). The pairwise backrub ensemble RMSDs continue to increase throughout the backrub simulation for all subsets (Fig. S6), indicating that diminishing returns at > 30,000 Monte Carlo steps is not a result of failure to sample new conformations, but rather might indicate that continued sampling does not capture additional relevant local changes in structure in this benchmark set.

Score analysis

As the sampling and scoring problems of protein modeling are generally linked, it is often the case that improving one enables further improvements in the other.

First, we compared the performance of our flex ddG method, which was run using Rosetta's Talaris [46,48,49] energy function, to an identical protocol run with the more recently developed Rosetta Energy Function (REF). [57] The REF energy function differs from the Talaris energy function by utilizing a new anisotropic implicit solvation model, and an improved electrostatics and Lennard-Jones model. REF was optimized simultaneously against small-molecule thermodynamic data and high-resolution macromolecular structural data. Using the REF energy function, we did not observe an increase in performance on the complete ZEMu dataset, and performance decreases were seen for the subsets of small-to-large mutations and multiple mutations (Table S8). Interestingly, flex ddG performance with the REF energy function increased over using the Talaris energy function if the resolution of the input crystal structure was $<=$ 1:5 Å, but this subset of the data was rather small with only 52 mutations.

Next, we sought to analyze underlying errors of the Rosetta energy function (when applied to interface $\Delta\Delta G$) by assessing the individual terms of the energy function. To do so, we chose to reweight the terms of the energy function using a non-linear reweighting scheme similar to Generalized Additive Models (GAMs). [52] In this reweighting method, we used Monte Carlo sampling to fit a sigmoid function to the individual distributions of energy function terms, with the objective function of reducing the absolute error between our predictions and known experimental values over the entire dataset.

The effect on the predictions is shown in Fig. 5, Fig. S7, and Table S9. In general, the GAM-adjusted predictions contain fewer outliers. In particular, experimental $\Delta\Delta G$ values that are relatively neutral (near zero) can sometimes be predicted by flex ddG to be highly destabilizing; the GAM model reduces the magnitude of error of many of these outliers, improving overall performance (Fig. 5). The overall correlation increases from 0.64 to 0.68 (Table 2 and Table S9) when refitting the values from the Rosetta Talaris energy function; [46,48,49] refitting values from the Rosetta REF energy function [57] leads to a similar increase from 0.63 to 0.68 (Fig. S7, Table S8, Table S9). The correlation coefficient also increases when refitting the values obtained for the no backrub control, but only to 0.62 (Fig. 5a, Table S9).

The fit functions (fit for Talaris-derived $\Delta\Delta G$ predictions) are shown in Fig. S8. Extreme values for most score terms are downweighted, especially for the fa_sol and fa_atr terms, which make the largest contributions to predicted $\Delta\Delta G$ (Fig. S9).

Conclusions

We have shown on a large, curated benchmark dataset that the “flex ddG” method presented here is more accurate than previous methods for estimating changes in binding affinity after mutation in protein-protein interfaces. Particular improvement in performance is seen on the subset of small-to-large mutations, indicating that representing backbone flexibility using backrub motions is effective in cases where backbone rearrangements are expected to be more common. Other notable improvements over previous methods are seen for stabilizing mutations, mutations in antibody-antigen interfaces, and for cases with multiple changes where none of the mutations is to an alanine residue.

We have also shown that more accurate predictions can be obtained by averaging the predictions across a generated structural ensemble of backrub models, and that the number of required models is relatively low (20–30). Prior methods that produced $\Delta\Delta G$ predictions by averaging an ensemble of models required on the order of thousands of models, [29] indicating that backrub sampling can efficiently sample the local conformational space around an input wild-type structure that is relevant for interface $\Delta\Delta G$ prediction.

By creating a method that uses backrub to sample conformational space more broadly than minimization alone, while still staying close to the known wild-type input structure, we have also generated data that should prove useful for future energy function improvements. In particular, using Rosetta’s newest REF energy function [57] does not improve performance of our method when compared to use of the prior Talaris [46,48,58] energy function (Table S8), indicating that the backrub sampling parameters might require further benchmarking and adaption to the REF energy function. Our error analysis via GAM-like reweighting also indicates potential avenues for energy function improvement by identifying imbalances in predicted energetic contributions leading to overestimation of stabilizing and destabilizing effects. Further improvements might also be obtained by more explicitly including the effects of altering water-mediated interactions [59] and of conformational entropy, [2,60] as well as by considering the commonly observed shortcomings of energy functions balancing the magnitudes of electrostatic interactions and desolvation costs. We expect energy function improvements to require more accurate representation of subtle conformational changes, as these changes can have a considerable impact on design predictions. [61]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge the following sources of funding: T.K. was supported by grants from the National Institute of Health (R01 GM110089 and R01 GM117189). M.H. was supported by Academy of Finland grant 299915. S.T. and J.E.L. were supported by National Science Foundation Graduate Research Fellowships. T.K. is a Chan Zuckerberg Biohub investigator.

References

1. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montaño B, Blundell TL, Ascher DB. Mutations at Protein-Protein Interfaces: Small Changes Over Big Surfaces Have Large Impacts on Human Health. Progress in Biophysics and Molecular Biology. 2017; 128:3–13. DOI: 10.1016/j.pbiomolbio.2016.10.002 [PubMed: 27913149]
2. Guerois R, Nielsen JE, Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. Journal of Molecular Biology. 2002; 320:369–387. DOI: 10.1016/S0022-2836(02)00442-4 [PubMed: 12079393]
3. Kamisetty H, Ramanathan A, Bailey-Kellogg C, Langmead CJ. Accounting for Conformational Entropy in Predicting Binding Free Energies of Protein-Protein Interactions. Proteins: Structure, Function, and Bioinformatics. 2011; 79:444–462. DOI: 10.1002/prot.22894
4. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: Prediction of Changes in Protein-Protein Binding Affinity on Mutations. Nucleic Acids Research. 2013; 41:W333–W339. DOI: 10.1093/nar/gkt450 [PubMed: 23723246]
5. Moal IH, Fernandez-Recio J. Intermolecular Contact Potentials for Protein-Protein Interactions Extracted From Binding Free Energy Changes Upon Mutation. Journal of Chemical Theory and Computation. 2013; 9:3715–3727. DOI: 10.1021/ct400295z [PubMed: 26584123]
6. Vangone A, Bonvin AM. Contacts-Based Prediction of Binding Affinity in Protein-Protein Complexes. eLife. 2015; 4:e07454.doi: 10.7554/eLife.07454 [PubMed: 26193119]
7. Brender JR, Zhang Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions Through Structure-Based Interface Profiles. PLOS Computational Biology. 2015; 11:e1004494.doi: 10.1371/journal.pcbi.1004494 [PubMed: 26506533]
8. Li M, Petukh M, Alexov E, Panchenko AR. Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. Journal of Chemical Theory and Computation. 2014; 10:1770–1780. DOI: 10.1021/ct401022c [PubMed: 24803870]
9. Tuncbag N, Gursoy A, Keskin O. Identification of Computational Hot Spots in Protein Interfaces: Combining Solvent Accessibility and Inter-Residue Potentials Improves the Accuracy. Bioinformatics. 2009; 25:1513–1520. DOI: 10.1093/bioinformatics/btp240 [PubMed: 19357097]

10. Pires DEV, Ascher DB, Blundell TL. mCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. Bioinformatics. 2014; 30:335–342. DOI: 10.1093/bioinformatics/btt691 [PubMed: 24281696]

11. Dourado DFAR, Flores SC. A Multiscale Approach to Predicting Affinity Changes in Protein-Protein Interfaces. Proteins: Structure, Function, and Bioinformatics. 2014; 82:2681–2690. DOI: 10.1002/prot.24634

12. Zhu X, Mitchell JC. KFC2: A Knowledge-Based Hot Spot Prediction Method Based on Interface Solvation, Atomic Density, and Plasticity Features. Proteins: Structure, Function, and Bioinformatics. 2011; 79:2671–2683. DOI: 10.1002/prot.23094

13. Mandell DJ, Kortemme T. Computer-Aided Design of Functional Protein Interactions. Nature Chemical Biology. 2009; 5:797–807. DOI: 10.1038/nchembio.251 [PubMed: 19841629]

14. Kaufmann KW, Lemmon GH, DeLuca SL, Sheehan JH, Meiler J. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. Biochemistry. 2010; 49:2987–2998. DOI: 10.1021/bi902153g [PubMed: 20235548]

15. Boulanger MJ, Bankovich AJ, Kortemme T, Baker D, Garcia KC. Convergent Mechanisms for Recognition of Divergent Cytokines by the Shared Signaling Receptor Gp130. Molecular Cell. 2003; 12:577–589. DOI: 10.1016/S1097-2765(03)00365-4 [PubMed: 14527405]

16. McFarland BJ, Kortemme T, Yu SF, Baker D, Strong RK. Symmetry Recognizing Asymmetry: Analysis of the Interactions Between the C-Type Lectin-like Immunoreceptor NKG2D and MHC Class I-like Ligands. Structure. 2003; 11:411–422. DOI: 10.1016/S0969-2126(03)00047-9 [PubMed: 12679019]

17. Sammond DW, Eletr ZM, Purbeck C, Kimple RJ, Siderovski DP, Kuhlman B. Structure-Based Protocol for Identifying Mutations That Enhance Protein-Protein Binding Affinities. Journal of Molecular Biology. 2007; 371:1392–1404. DOI: 10.1016/j.jmb.2007.05.096 [PubMed: 17603074]

18. Song G, Lazar GA, Kortemme T, Shimaoka M, Desjarlais JR, Baker D, Springer TA. Rational Design of Intercellular Adhesion Molecule-1 (ICAM-1) Variants for Antagonizing Integrin Lymphocyte Function-Associated Antigen-1-Dependent Adhesion. Journal of Biological Chemistry. 2006; 281:5042–5049. DOI: 10.1074/jbc.M510454200 [PubMed: 16354667]

19. Kapp GT, Liu S, Stein A, Wong DT, Reményi A, Yeh BJ, Fraser JS, Taunton J, Lim WA, Kortemme T. Control of Protein Signaling Using a Computationally Designed GTPase/GEF Orthogonal Pair. Proceedings of the National Academy of Sciences. 2012; 109:5277–5282. DOI: 10.1073/pnas.1114487109

20. Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ, Stoddard BL. Design, Activity, and Structure of a Highly Specific Artificial Endonuclease. Molecular Cell. 2002; 10:895–905. DOI: 10.1016/S1097-2765(02)00690-1 [PubMed: 12419232]

21. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. Science. 2011; 332:816–821. DOI: 10.1126/science.1202617 [PubMed: 21566186]

22. Chevalier A, Silva DA, Rocklin GJ, Hicks DR, Vergara R, Murapa P, Bernard SM, Zhang L, Lam KH, Yao G, et al. Massively Parallel De Novo Protein Design for Targeted Therapeutics. Nature. 2017; 550:74–79. DOI: 10.1038/nature23912 [PubMed: 28953867]

23. Kellogg EH, Leaver-Fay A, Baker D. Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability. Proteins: Structure, Function, and Bioinformatics. 2011; 79:830–838. DOI: 10.1002/prot.22921

24. Kortemme T, Baker D. A Simple Physical Model for Binding Energy Hot Spots in Protein-Protein Complexes. Proceedings of the National Academy of Sciences. 2002; 99:14116–14121. DOI: 10.1073/pnas.202485799

25. Kortemme T, Kim DE, Baker D. Computational Alanine Scanning of Protein-Protein Interfaces. Science Signaling. 2004; 2004:pl2–pl2. DOI: 10.1126/stke.2192004pl2

26. Conchúir SÓ, Barlow KA, Pache RA, Ollikainen N, Kundert K, O'Meara MJ, Smith CA, Kortemme T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. PLOS ONE. 2015; 10:e0130433.doi: 10.1371/journal.pone.0130433 [PubMed: 26335248]

27. Cunningham BC, Wells JA. High-Resolution Epitope Mapping of hGH-receptor Interactions by Alanine-Scanning Mutagenesis. Science. 1989; 244:1081–1085. DOI: 10.1126/science.2471267 [PubMed: 2471267]

28. Davey JA, Damry AM, Euler CK, Goto NK, Chica RA. Prediction of Stable Globular Proteins Using Negative Design with Non-native Backbone Ensembles. Structure. 2015; 23:2011–2021. DOI: 10.1016/j.str.2015.07.021 [PubMed: 26412333]

29. Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. Predicting Free Energy Changes Using Structural Ensembles. Nature Methods. 2009; 6:3–4. DOI: 10.1038/nmeth0109-3 [PubMed: 19116609]

30. Araki M, Kamiya N, Sato M, Nakatsui M, Hirokawa T, Okuno Y. The Effect of Conformational Flexibility on Binding Free Energy Estimation Between Kinases and Their Inhibitors. Journal of Chemical Information and Modeling. 2016; 56:2445–2456. DOI: 10.1021/acs.jcim.6b00398 [PubMed: 28024406]

31. Smith CA, Kortemme T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. Journal of Molecular Biology. 2008; 380:742–756. DOI: 10.1016/j.jmb.2008.05.023 [PubMed: 18547585]

32. Davis IW, Arendall WB, Richardson DC, Richardson JS. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. Structure. 2006; 14:265–274. DOI: 10.1016/j.str.2005.10.007 [PubMed: 16472746]

33. Keedy DA, Georgiev I, Triplett EB, Donald BR, Richardson DC, Richardson JS. The Role of Local Backrub Motions in Evolved and Designed Mutations. PLOS Computational Biology. 2012; 8:e1002629.doi: 10.1371/journal.pcbi.1002629 [PubMed: 22876172]

34. Friedland GD, Linares AJ, Smith CA, Kortemme T. A Simple Model of Backbone Flexibility Improves Modeling of Side-Chain Conformational Variability. Journal of Molecular Biology. 2008; 380:757–774. DOI: 10.1016/j.jmb.2008.05.006 [PubMed: 18547586]

35. Humphris EL, Kortemme T. Prediction of Protein-Protein Interface Sequence Diversity Using Flexible Backbone Computational Protein Design. Structure. 2008; 16:1777–1788. DOI: 10.1016/j.str.2008.09.012 [PubMed: 19081054]

36. Smith CA, Kortemme T. Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. Journal of Molecular Biology. 2010; 402:460–474. DOI: 10.1016/j.jmb.2010.07.032 [PubMed: 20654621]

37. Smith CA, Kortemme T. Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. PLOS ONE. 2011; 6:e20451.doi: 10.1371/journal.pone.0020451 [PubMed: 21789164]

38. Schenkelberg CD, Bystroff C. Protein Backbone Ensemble Generation Explores the Local Structural Space of Unseen Natural Homologs. Bioinformatics. 2016; 32:1454–1461. DOI: 10.1093/bioinformatics/btw001 [PubMed: 26787668]

39. Ollikainen N, Jong RMd, Kortemme T. Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-Design of Protein-Ligand Specificity. PLOS Comput Biol. 2015; 11:e1004335.doi: 10.1371/journal.pcbi.1004335 [PubMed: 26397464]

40. Davey JA, Chica RA. Improving the Accuracy of Protein Stability Predictions With Multistate Design Using a Variety of Backbone Ensembles. Proteins: Structure, Function, and Bioinformatics. 2014; 82:771–784. DOI: 10.1002/prot.24457

41. Moal IH, Fernández-Recio J. SKEMPI: A Structural Kinetic and Energetic Database of Mutant Protein Interactions and Its Use in Empirical Models. Bioinformatics. 2012; 28:2600–2607. DOI: 10.1093/bioinformatics/bts489 [PubMed: 22859501]

42. Geng C, Vangone A, Bonvin AMJJ. Exploring the Interplay Between Experimental Methods and the Performance of Predictors of Binding Affinity Change Upon Mutations in Protein Complexes. Protein Engineering, Design and Selection. 2016; 29:291–299. DOI: 10.1093/protein/gzw020

43. Simpson, RJ. Proteins and Proteomics: A Laboratory Manual. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 2002. lab manual edition

44. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. SAbDab: The Structural Antibody Database. Nucleic Acids Research. 2014; 42:D1140–D1146. DOI: 10.1093/nar/gkt1043 [PubMed: 24214988]

45. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G, et al. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. PLOS ONE. 2011; 6:e20161.doi: 10.1371/journal.pone.0020161 [PubMed: 21731610]

46. Shapovalov MV, Dunbrack RL Jr. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived From Adaptive Kernel Density Estimates and Regressions. Structure. 2011; 19:844–858. DOI: 10.1016/j.str.2011.03.019 [PubMed: 21645855]

47. Leaver-Fay A, Jacak R, Stranges PB, Kuhlman B. A Generic Program for Multi-state Protein Design. PLOS ONE. 2011; 6:e20937.doi: 10.1371/journal.pone.0020937 [PubMed: 21754981]

48. Song Y, Tyka M, Leaver-Fay A, Thompson J, Baker D. Structure-Guided Force-field Optimization. Proteins: Structure, Function, and Bioinformatics. 2011; 79:1898–1909. DOI: 10.1002/prot.23013

49. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, DiMaio F, Bradley P, Kortemme T, Baker D, Snoeyink J, et al. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction With Rosetta. Journal of Chemical Theory and Computation. 2015; 11:609–622. DOI: 10.1021/ct500864r [PubMed: 25866491]

50. Lazaridis T, Karplus M. Effective Energy Function for Proteins in Solution. Proteins: Structure, Function, and Bioinformatics. 1999; 35:133–152.

51. Kortemme T, Morozov AV, Baker D. An Orientation-Dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes. Journal of Molecular Biology. 2003; 326:1239–1259. DOI: 10.1016/S0022-2836(03)00021-4 [PubMed: 12589766]

52. Hastie, TJ., Tibshirani, RJ. Generalized Additive Models. 1. Chapman and Hall/CRC; Boca Raton, Fla: 1990.

53. Howell SC, Inampudi KK, Bean DP, Wilson CJ. Understanding Thermal Adaptation of Enzymes through the Multistate Rational Design and Stability Prediction of 100 Adenylate Kinases. Structure. 2014; 22:218–229. DOI: 10.1016/j.str.2013.10.019 [PubMed: 24361272]

54. Davey JA, Chica RA. Optimization of Rotamers Prior to Template Minimization Improves Stability Predictions Made by Computational Protein Design. Protein Science. 2015; 24:545–560. DOI: 10.1002/pro.2618 [PubMed: 25492709]

55. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers. 1983; 22:2577–2637. DOI: 10.1002/bip.360221211 [PubMed: 6667333]

56. Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G. A Series of PDB Related Databases for Everyday Needs. Nucleic Acids Research. 2011; 39:D411–D419. DOI: 10.1093/nar/gkq1105 [PubMed: 21071423]

57. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. Journal of Chemical Theory and Computation. 2017; 13:3031–3048. DOI: 10.1021/acs.jctc.7b00125 [PubMed: 28430426]

58. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, et al. Chapter Six - Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. Methods in Enzymology. 2013; 523:109–143. DOI: 10.1016/B978-0-12-394292-0.00006-0 [PubMed: 23422428]

59. Lai JK, Ambia J, Wang Y, Barth P. Enhancing Structure Prediction and Design of Soluble and Membrane Proteins With Explicit Solvent-Protein Interactions. Structure. 2017; 25:1758–1770e8. DOI: 10.1016/j.str.2017.09.002 [PubMed: 28966016]

60. Hu X, Kuhlman B. Protein Design Simulations Suggest That Side-Chain Conformational Entropy Is Not a Strong Determinant of Amino Acid Environmental Preferences. Proteins: Structure, Function, and Bioinformatics. 2006; 62:739–748. DOI: 10.1002/prot.20786

61. Dou J, Doyle L Jr, Greisen P, Schena A, Park H, Johnsson K, Stoddard BL, Baker D. Sampling and Energy Evaluation Challenges in Ligand Binding Protein Design. Protein Science. 2017; 26:2426–2437. DOI: 10.1002/pro.3317 [PubMed: 28980354]

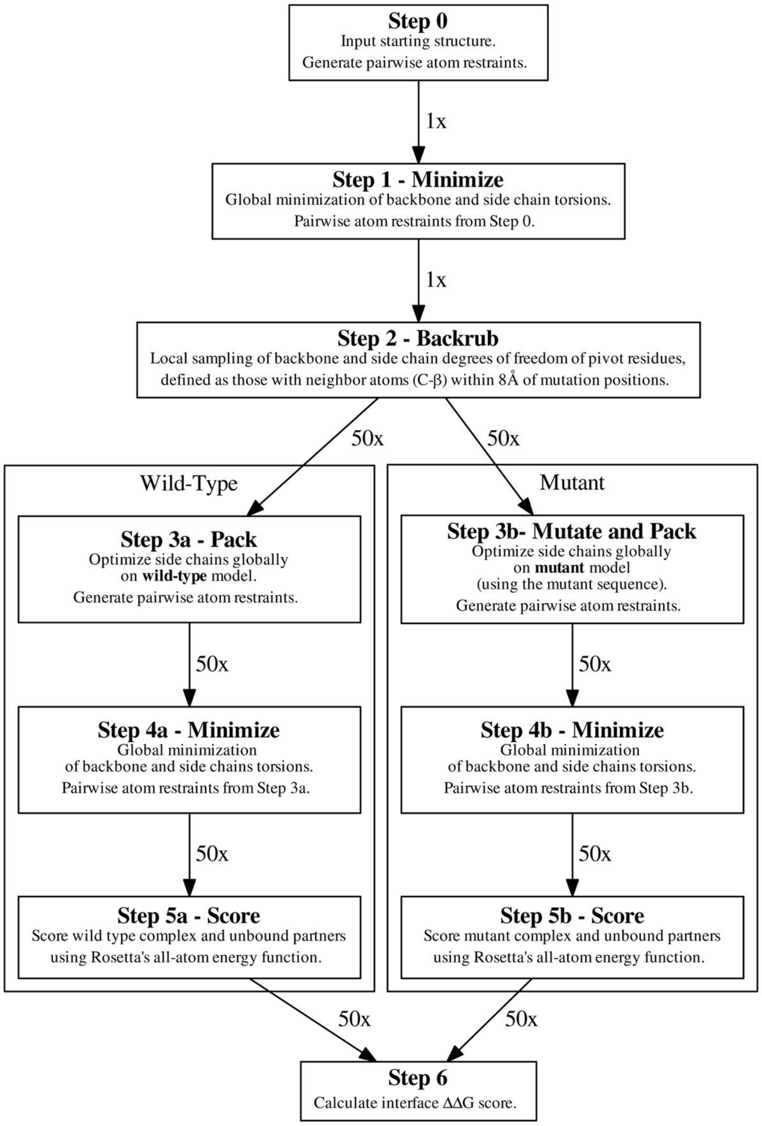

Step 0
Input starting structure.
Generate pairwise atom restraints.
1x
Step 1 - Minimize
Global minimization of backbone and side chain torsions.
Pairwise atom restraints from Step 0.
1x
Step 2 - Backrub
Local sampling of backbone and side chain degrees of freedom of pivot residues,
defined as those with neighbor atoms (C-β) within 8Å of mutation positions.
50x
50x
Wild-Type
Mutant
Step 3a - Pack
Optimize side chains globally
on wild-type model.
Generate pairwise atom restraints.
Step 3b- Mutate and Pack
Optimize side chains globally
on mutant model
(using the mutant sequence).
Generate pairwise atom restraints.
50x
50x
Step 4a - Minimize
Global minimization
of backbone and side chains torsions.
Pairwise atom restraints from Step 3a.
Step 4b - Minimize
Global minimization
of backbone and side chains torsions.
Pairwise atom restraints from Step 3b.
50x
50x
Step 5a - Score
Score wild type complex and unbound partners
using Rosetta's all-atom energy function.
Step 5b - Score
Score mutant complex and unbound partners
using Rosetta's all-atom energy function.
50x
50x
Step 6
Calculate interface ΔΔG score.

Figure 1.

Schematic of the flex ddG protocol method.

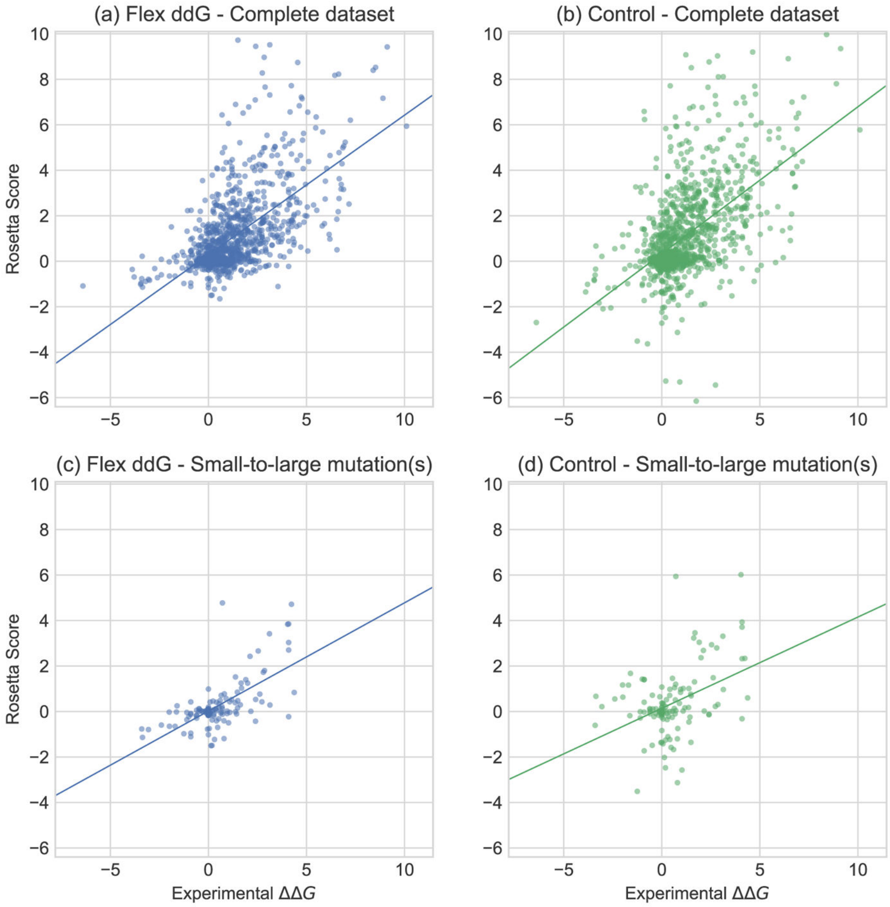


Figure 2.

Experimentally determined ΔΔ*G* values (x-axis) versus Rosetta predictions. Rosetta scores are in Rosetta Energy Units (REU) using the Rosetta Talaris energy function. [46,48,49] (a) flex ddG method (35000 backrub steps); Complete dataset (n=1240). (b) no backrub control; Complete dataset (n=1240). (c) flex ddG method (35000 backrub steps); Small-to-large mutation(s) (n=130). (d) no backrub control; Small-to-large mutation(s) (n=130).

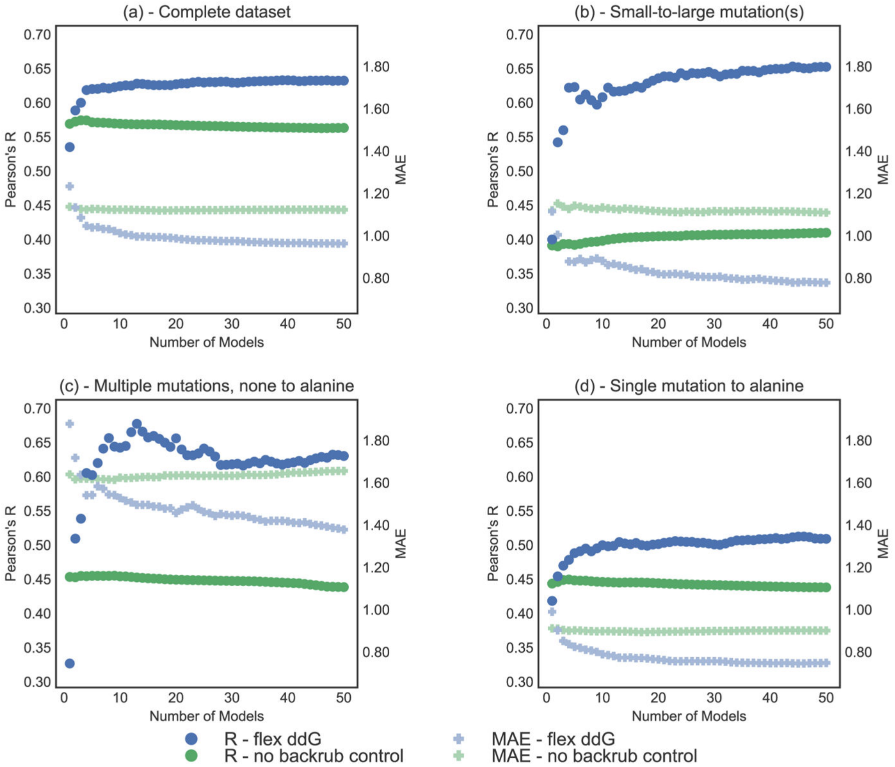


Figure 3.

Correlation (Pearson's R, left y-axis) and MAE (Mean Absolute Error, right y-axis) vs. number of averaged models (x-axis), on the complete ZEMu set, and subsets. Pearson's R is shown as circles, and MAE as faded plusses. Predictions generated with the Flex ddG protocol are shown in blue. Predictions generated with the no backrub control protocol are shown in green. A selection of key data underlying this figure can be found in Table S4. Flex ddG is run with 35000 backrub steps. Structures are sorted by their minimized wild-type complex energy. (a) Complete dataset (n = 1240) (b) Small-to-large mutation(s) (n = 130) (c) Multiple mutations, none to alanine (n = 45) (d) Single mutation to alanine (n = 748).

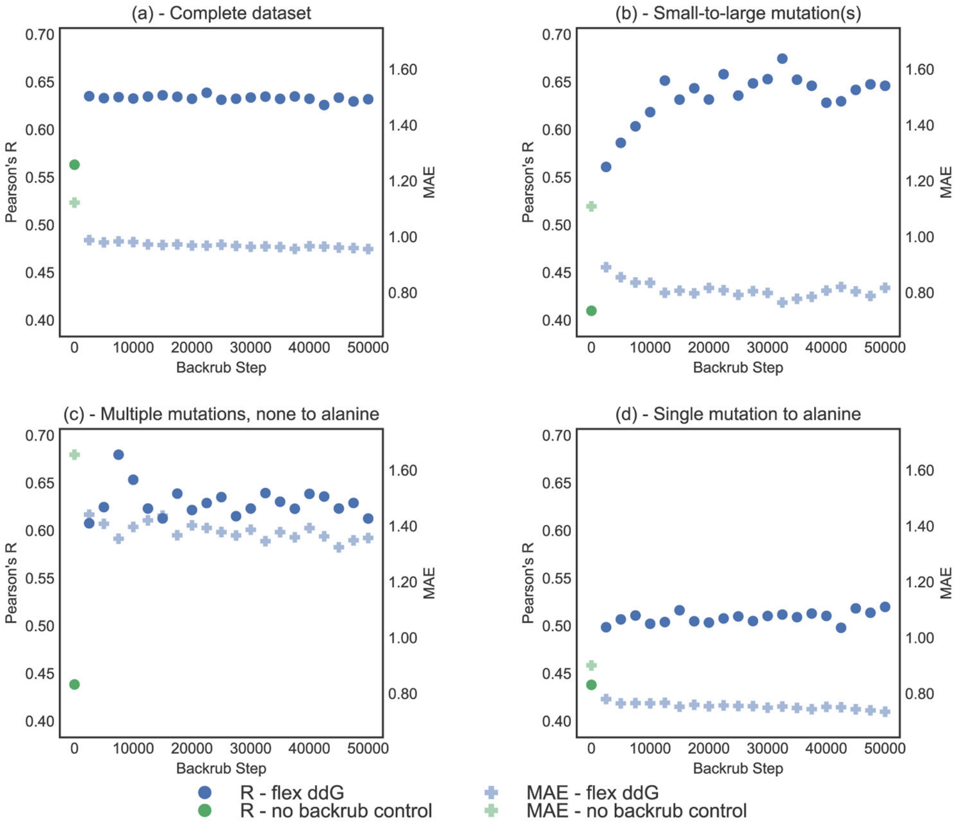


Figure 4.

Correlation (Pearson's R) and MAE (Mean Absolute Error) vs. number of backrub steps, on the complete ZEMu set, and subsets. Pearson's R is shown as circles, and MAE as faded plusses. Predictions generated with the Flex ddG protocol are shown in blue. Predictions generated with the no backrub control protocol are shown in green. A selection of key data underlying this figure can be found in Table S7. (a) Complete dataset (n=1240) (b) Small-to-large mutation(s) (n=130) (c) Multiple mutations, none to alanine (n=45) (d) Single mutation to alanine (n=748)

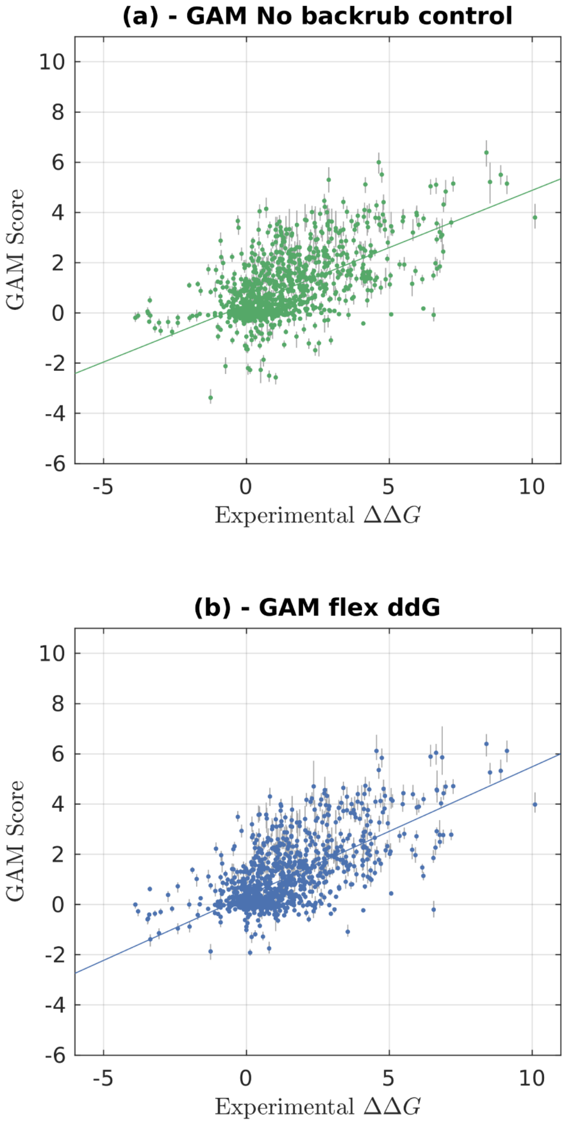


Figure 5.

Experimentally determined $\Delta\Delta G$ values (x-axis) versus predictions using a Generalized additive model (GAM). The complete dataset is shown. GAM scores are refit from values in Rosetta Energy Units (REU) using the Rosetta Talaris [46,48,49] energy function. The error bars in gray represent the range from minimum to maximum fit predicted $\Delta\Delta G$ value for the 1000 sampled GAM models. **(a)**: Control (no backrub) Rosetta predictions. **(b)**: Flex ddG Rosetta predictions using 35,000 backrub steps and 50 output models. A line of best fit is shown in each of the panels.

Table 1

ZEMu dataset composition

n	Name
1240	Complete dataset
748	Single mutation to alanine
273	Multiple mutations
130	Small-to-large mutation(s)
45	Multiple mutations, none to alanine

Table 2

Summary of prediction performance. Flex ddG predictions used 50 models and 35000 backrub steps. ddG monomer predictions used the default of averaging the $\Delta\Delta G$ scores of the three lowest scoring output models, as implemented in the original method. [23] N = number of cases in the dataset or subset. R = Pearson’s R. MAE = Mean Absolute Error. FC = Fraction Correct. Best performance for each metric and dataset is shown in bold.

Mutation Category	Prediction Method	N	R	MAE	FC
Complete dataset	flex ddG	1240	**0.63**	**0.96**	**0.76**
	ddG monomer		0.51	1.57	0.64
	no backrub control		0.56	1.12	0.73
	ZEMu paper		0.61	1.08	0.71
Small-to-large mutation(s)	flex ddG	130	**0.65**	**0.78**	**0.71**
	ddG monomer		0.31	1.55	0.55
	no backrub control		0.41	1.11	0.62
	ZEMu paper		0.48	1.16	0.65
Mutation(s) to alanine	flex ddG	939	**0.62**	**0.96**	**0.78**
	ddG monomer		0.50	1.55	0.66
	no backrub control		0.58	1.06	0.75
	ZEMu paper		**0.62**	1.03	0.73
Single mutation to alanine	flex ddG	748	**0.51**	**0.75**	**0.76**
	ddG monomer		0.36	1.31	0.62
	no backrub control		0.44	0.90	0.74
	ZEMu paper		0.45	0.86	0.71
Multiple mutations	flex ddG	273	0.62	**1.62**	**0.78**
	ddG monomer		0.50	2.44	0.70
	no backrub control		0.58	1.73	0.73
	ZEMu paper		**0.64**	1.63	0.75
Multiple mutations, all to alanine	flex ddG	191	0.47	1.77	**0.84**
	ddG monomer		0.34	2.49	0.80
	no backrub control		0.50	**1.69**	0.81
	ZEMu paper		**0.55**	1.72	0.79

Mutation Category	**Prediction Method**	**N**	**R**	**MAE**	**FC**
Multiple mutations, none to alanine	flex ddG	45	**0.63**	**1.38**	**0.60**
	ddG monomer		0.40	2.54	0.38
	no backrub control		0.44	1.66	0.58
	ZEMu paper		0.53	1.59	**0.60**
Antibodies	flex ddG	355	**0.61**	**0.93**	**0.74**
	ddG monomer		0.50	1.35	0.69
	no backrub control		0.49	1.06	0.72
	ZEMu paper		0.54	1.06	0.67