Quantification of biases in predictions of protein-protein binding affinity changes upon mutations

Matsvei Tsishyn[1,2,*], Fabrizio Pucci[1,2,†,*], Marianne Rooman[1,2,†,*]

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles, Bruxelles, Belgium.

[2] Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium.

[†] Contributed equally to this work.

[*] Correspondence at matsvei.tsishyn@ulb.be, fabrizio.pucci@ulb.be, marianne.rooman@ulb.be

August 4, 2023

Abstract

Understanding the impact of mutations on protein-protein binding affinity is a key objective for a wide range of biotechnological applications and for shedding light on disease-causing mutations, which are often located at protein-protein interfaces. Over the past decade, many computational methods using physics-based and/or machine learning approaches have been developed to predict how protein binding affinity changes upon mutations. They all claim to achieve astonishing accuracy on both training and test sets, with performances on standard benchmarks such as SKEMPI 2.0 that seem overly optimistic. Here we benchmarked six well-known and well-used predictors and identified their biases and dataset dependencies, using not only SKEMPI 2.0 as a test set but also deep mutagenesis data on the SARS-CoV-2 spike protein in complex with the human angiotensin-converting enzyme 2. We showed that, even though most tested methods reach a significant degree of robustness and accuracy, they suffer from limited generalizability properties and struggle to predict unseen mutations. Undesirable prediction biases towards specific mutation properties, the most marked being towards destabilizing mutations, are also observed and should be carefully considered by method developers. We conclude from our analyses that there is room for improvement of the prediction models and propose ways to check, assess and improve their generalizability and robustness.

1 Introduction

Proteins interact with each other and form complexes that perform a wide range of biological functions in the intra- and extracellular media, involved in key processes such as signal transduction, cell growth and proliferation, and cell apoptosis. It is

therefore of fundamental interest to understand how amino acid substitutions impact on the ability of proteins to bind to their interacting partners. Such insights would shed light on pathogenic mechanisms since aberrant protein-protein interactions caused by deleterious variants are often central to Mendelian disorders and complex diseases such as cancer [1, 2]. From a biotechnological perspective, it would improve the design of drugs that modulate protein-protein interactions (PPI), as targeting these is an established strategy in the treatment of diseases [3, 4].

There are several experimental methods to estimate the impact of mutations on PPIs. Biophysical methods such as isothermal titration calorimetry allow in-depth estimation of protein binding thermodynamics [5]; in contrast, high-throughput screening assays such as yeast-two-hybrid systems only allow identification of binary PPIs but have the advantage of being applicable at large scale [6]. However, given that all experimental approaches remain challenging, costly and time-intensive, there is room for computational methods which provide effective alternatives to predict and achieve better understanding of PPIs.

Many studies have been dedicated over the last decade to the development of bioinformatics tools to predict the impact of mutations on protein-protein binding affinity (ΔG_b), which is the thermodynamic descriptor of protein-protein interactions [7, 8, 9, 10, 11, 12]. These tools are mainly based on structural features derived from experimentally characterized protein complexes and/or evolutionary data. These features are usually combined using standard machine learning techniques, but deep learning algorithms are starting to be used in predictor construction [12]. The different tools are generally trained and tested on sets of experimental values of the change in protein-protein binding affinity upon mutations ($\Delta\Delta G_b$), which are manually collected from the literature [13, 14]. More on the development of $\Delta\Delta G_b$ predictions and experimental datasets is exposed in Supplementary Section 1.

While these tools reach good prediction accuracy on their respective training sets, the extent to which these results are generalizable to unseen data is one of the open issues in the field. Indeed, like all supervised machine learning methods, they are likely to suffer from undesirable biases towards the learning set, which often hinder the generalization of their predictions. One example of this problem is the bias towards destabilizing values of the change in folding free energy upon mutations ($\Delta\Delta G$), which has been thoroughly analyzed in a series of investigations [15, 16, 17]. In summary, it has been shown that training protein stability predictors on the common experimental datasets that are dominated by destabilizing mutations lead to much better performance on destabilizing than on stabilizing mutations.

Although prediction biases have been studied for predictors of stability changes caused by mutations, they have not been for protein-protein affinity changes; yet having accurate and unbiased prediction tools of $\Delta\Delta G_b$ values is crucial for a wide range of biotechnological applications. In this paper, we systematically quantified possible biases in state-of-the-art protein-protein $\Delta\Delta G_b$ prediction methods. More precisely, we evaluated their predictions on a set of mutations with experimentally measured $\Delta\Delta G_b$ values taken from [14], and on high-throughput data on the binding between the human angiotensin-converting enzyme 2 (ACE2) and the receptor binding domain (RBD) of the SARS-CoV-2 spike protein taken from [18]. After an analysis of the methods' performances, we suggest strategies to limit and correct possible biases and thus to further improve the methods' generalizability and scores.

2 Methods

2.1 Protein-protein binding affinity change upon mutations

The thermodynamic protein-protein binding affinity ΔG_b is a measure of the strength of a PPI and is defined using the Gibbs free energy:

$$\Delta G_b := RT \ln K_D, \tag{1}$$

where R is the Boltzmann constant, T the absolute temperature (in K) and K_D the equilibrium dissociation constant of the PPI. We use the convention that the stronger the interaction, the more negative the ΔG_b value, and express it in kcal/mol.

Under the action of a mutation, we define the binding affinity change as:

$$\Delta\Delta G_b^{wt\to mt} := \Delta G_b^{mt} - \Delta G_b^{wt} = RT \ln \frac{K_D^{mt}}{K_D^{wt}}. \tag{2}$$

where wt refers to the wild-type complex and mt to the mutant. Thus, positive $\Delta\Delta G_b$ values correspond to mutations that destabilize the complex and negative values to stabilizing mutations. Since binding affinity is a thermodynamic state function, mutating from a wild-type complex to a mutant complex and then mutating back results in no change in $\Delta\Delta G_b$, which is expressed by the following equation:

$$\Delta\Delta G_b^{wt\to mt} + \Delta\Delta G_b^{mt\to wt} = 0. \tag{3}$$

We will refer to this property as the symmetry property.

In what follows, we call "direct mutation" a mutation that goes from the wild-type to the mutant complex. Conversely, we call "reverse mutation" a mutation that goes from the mutant to the wild-type complex. Note that the terms wild-type, mutant, direct and reverse are defined with respect to the proteins that are part of our datasets and do not necessarily have any biological interpretation.

2.2 Defining protein-protein interfaces

The relative solvent accessibility (RSA) of a residue in a 3-dimensional (3D) structure is defined as the ratio (in %) of its solvent-accessible surface area in the structure and in an extended tripeptide Gly-X-Gly [19]. We computed it using our in-house software MuSiC [20] (which uses the DSSP algorithm [21]) available on the dezyme.com website. We distinguish between interactant-RSA (iRSA) and complex-RSA (cRSA), which correspond to the RSA computed from the structure containing solely the considered interactant and from the structure containing the complex with both interactants, respectively. We defined the RSA change upon binding as $\Delta\text{RSA} := \text{iRSA} - \text{cRSA}$. Finally, a residue is considered to be in the protein-protein interface if its ΔRSA is greater than 5%.

2.3 Datasets of binding affinity changes upon mutations

We considered two datasets. The first is based on the SKEMPI sets [13, 14], containing mutations in different protein-protein complexes of known 3D structure available

in the Protein Data Bank (PDB) [22], whose $\Delta\Delta G_b$ values were experimentally measured using biophysical methods, performed by various laboratories. The number of characterized mutations in each protein typically ranges from a few to a few dozen, and reaches a few hundred in some rare cases [23, 24]. These datasets yield relatively accurate $\Delta\Delta G_b$ values but have the disadvantages of being non-systematic and of reflecting the specific interests of the authors in the choice of proteins and mutations.

The SKEMPI 2.0 dataset [14], contains 7085 entries and is the most comprehensive, well-curated and diverse dataset of its kind. First we discarded entries without $\Delta\Delta G_b$ value and entries describing multiple mutations. Then we aggregated all redundant entries (with the same mutation in the same PDB structure) by taking their average $\Delta\Delta G_b$ value. To withdraw the dependency on the quality of the structures, we also dropped all mutations in low-resolution X-ray structures (resolution > 2.5Å) and in structures obtained by nuclear magnetic resonance spectroscopy. This defines our first benchmark dataset called S2536 that contains 2536 mutations in 205 different PDB structures. We used the 3D structures of the protein complexes that were collected in the SKEMPI 2.0 database, as they were curated to be as close as possible to the protein complexes on which the measurements were made.

The second dataset we considered contains affinity values obtained through deep mutagenesis experiments that systematically characterized all possible mutations in the receptor binding domain (RBD) of the SARS-CoV-2 spike glycoprotein in interaction with the human ACE2 receptor [18]. The experimental 3D structure of the ACE2-RBD complex in which all these mutations were inserted has the PDB ID 6M0J [25]. This dataset has the advantage of being systematic and thus less biased. However, the measured values are not exact $\Delta\Delta G_b$ but close correlates. From this set, we first discarded the mutations of the few residues located in the N- and C-terminal tails of the spike protein, as they are absent from the reference PDB structure 6M0J. Then we identified the residues of the ACE2-RBD interface using the RSA criterion specified above, of which there are 20. We focused on all 380 possible mutations of these 20 residues, to define our second benchmark dataset C380.

For both datasets S2536 and C380, considered by definition as direct mutations, we constructed the datasets of reverse mutations using the symmetry property Eq. (3) to assign a $\Delta\Delta G_b$ value to each reverse mutation. The structures of the mutant complexes were modeled using the comparative modeling software MODELLER [26] with default parameters and using the wild-type structures as templates. When the distinction is required, we append the suffix -D to the name for a dataset of direct mutations, the suffix -R for a dataset of reverse mutations and the suffix -DR for a dataset of both direct and reverse mutations (e.g., S2536-D, S2536-R and S2536-DR). The datasets S2536 and C380 are available at https://github.com/3BioCompBio/DDGb_bias and all wild-type (experimental) and mutant (modeled) structures can be downloaded at http://babylone.3bio.ulb.ac.be/DDGb_bias_structures/.

2.4 Prediction methods tested

We benchmarked the six best known, available, and widely used $\Delta\Delta G_b$ predictors published in recent years. We briefly describe their properties.

mCSM-PPI2 [7] is a machine learning predictor that uses graph-based structural signatures of the inter-residue interaction network, evolutionary information, complex network metrics and energy terms.

MutaBind2 [8] uses seven features among which protein-solvent interactions, evolutionary conservation and physics-based thermodynamic stability.

BeAtMuSiC [9] is our in-house predictor. It estimates $\Delta\Delta G_b$ through a linear combination of the stability changes upon mutations computed by the PoPMuSiC predictor [27] of the protein complex and the separate interactants.

SSIPe [10] combines protein interface profiles collected from structure and sequence homology searches with physics-based energy functions.

SAAMBE-3D [11] is a machine learning-based predictor that utilizes 33 knowledge-based features representing the physical environment surrounding the mutation site.

NetTree [12] is a deep-learning method based on convolutional neural networks and algebraic topology features. It uses element- and site-specific persistent homology to simplify the structure of a protein complex and to convert it into topological features.

These predictors can be classified into three groups based on their type of approach: mCSM-PPI2, MutaBind2 and SSIPe use evolutionary and physics-based information that are combined using machine learning techniques; SAAMBE-3D and NetTree are pure machine learning predictors whose features are extracted from protein structures; and BeAtMuSiC exploits a physics-based stability predictor.

In terms of training set, we have the following classification: NetTree was trained on antigen-antibody interaction data from the AB-Bind dataset [28]; BeAtMuSiC was released before SKEMPI 2.0 and thus trained on SKEMPI 1.0 entries; all other predictors were trained on entries from SKEMPI 2.0; mCSM-PPI2 and MutaBind2 included, in addition to direct mutations, reverse mutations in their training datasets.

Predictions from BeAtMuSiC, SSIPe, SAAMBE-3D and NetTree were obtained by executing their standalone code, while predictions from mCSM-PPI2 and MutaBind2 were obtained using their online webserver.

3 Results and discussion

3.1 An upper bound to the accuracy of predictors

Values of binding affinity changes collected from the literature, available in S2536, come from experiments performed using different techniques and in different environmental conditions such as pH, temperature or solvent additives. These differences add to the experimental errors and usually lead to different $\Delta\Delta G_b$ values for the same mutation in the same protein complex. Moreover, even if SKEMPI 2.0 is particularly well curated, curation errors cannot be avoided, as illustrated by the error corrections between SKEMPI 1.0 and SKEMPI 2.0 (see Supplementary Section 2). The uncertainty on $\Delta\Delta G_b$ values creates an upper bound to the precision of the predictions, which cannot surpass the accuracy of experimental data.

An analytical method to estimate the upper bound on the Pearson correlation coefficient (ρ) and the lower bound on the root mean squared error (RMSE) was recently proposed [29, 30]. These bounds are expressed as:

$$\sup(\rho) = \frac{\sigma_{DB}^2}{\bar{\sigma}^2 + \sigma_{DB}^2}, \tag{4}$$

$$\inf(\text{RMSE}) = \sqrt{2\bar{\sigma}^2}, \tag{5}$$

where σ^2_{DB} is the variance of $\Delta\Delta G_b$ values in the whole dataset and $\bar{\sigma}^2$ is the mean of individual variances for redundant entries. We estimated the values of these bounds using the 116 redundant clusters with at least three entries among all single mutations from the SKEMPI 2.0 dataset.

We obtained: $\sup(\rho) = 0.89$ and $\inf(\text{RMSE}) = 0.70$ kcal/mol. Note, however, that these bounds are probably overestimated and underestimated, respectively, due to a underestimation of $\bar{\sigma}^2$. Indeed, only independent, uncorrelated, $\Delta\Delta G_b$ measures of a given mutation can yield a correct estimation of the variance, which seems to not be always the case.

The performances of the tested predictors presented in the following sections can be compared to these "optimal" values. It has to be emphasized that any accuracy better than these bounds can only be explained by overfitting.

3.2 Biases in the S2536 dataset

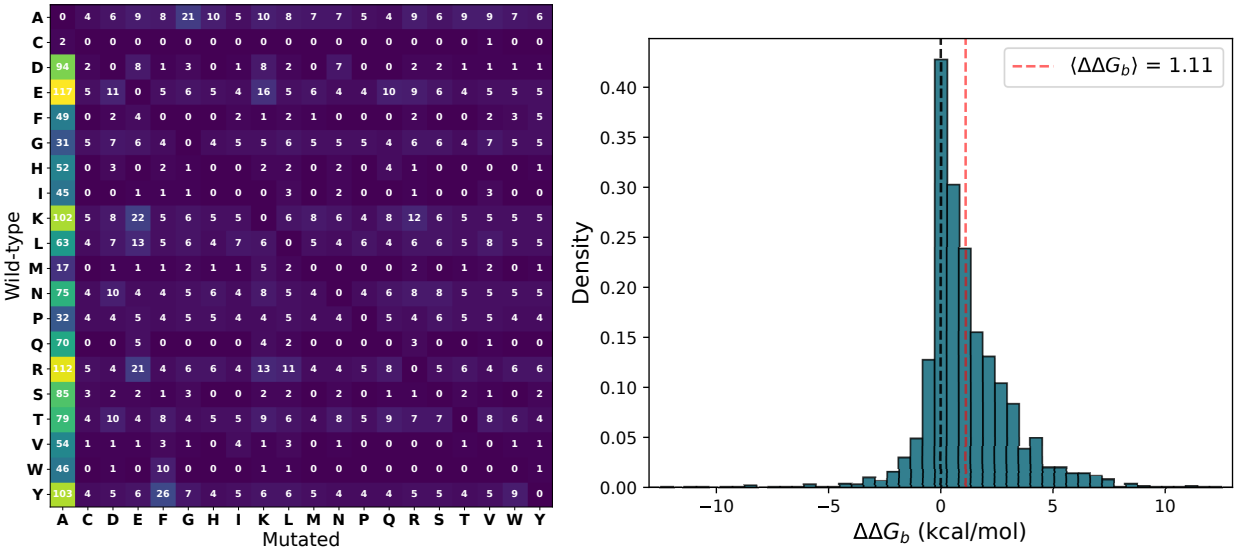
As mentioned by SKEMPI authors [13, 14], mutations characterized and reported in the literature are not systematic but reflect the interests of the experimenters. The collected data have thus biases towards specific residues, mutation types, spatial locations, proteins and protein families. These biases can lead to overoptimistic evaluations of the predictors, even if strict cross-validation methods are used. Indeed, when training and test sets are submitted to the same biases, a predictor can learn and replicate them, which increases both its apparent performance and generalization error. This may create a gap between the performances estimated from either a biased test set or a set of systematic mutations, and thus raises concerns about predictors reliability. In this section we quantified and discussed some of the biases in the S2536 mutations set.

First we note the imbalance in terms of mutation types. The number of occurrences of the 380 possible mutation types in S2536 are shown in Fig. 1a. Half of the mutations are towards alanine, 222 mutation types occur less than 5 times and 92 mutation types are unrepresented. This tendency is related to the prevalence of experimental alanine-scanning data in S2536. It may weaken predictions of underrepresented mutation types.

Another notable imbalance is towards mutations located at protein-protein interfaces: 78% of S2536 entries are mutations of the 9% residues located at the interface. Even though interface residues are usually more critical for the interaction, non-interface regions can also be important and their impact risks being overlooked by predictors.

Finally, the $\Delta\Delta G_b$ distribution is largely shifted towards positive values, as shown in Fig. 1b. It has a mean value of 1.11 kcal/mol and a standard deviation of 1.99 kcal/mol with clear prevalence of destabilizing mutations. This imbalance is not surprising since experimentally studied complexes are often optimized for high binding affinity by evolution. However, it tends to cause predictors to systematically output destabilizing $\Delta\Delta G_b$ values even for neutral and stabilizing mutations, thus preventing to satisfy the symmetry property (Eq. (3)). This issue, which is particularly problematic for, e.g., rational protein design, has been identified and widely investigated in the context of stability changes upon mutations [17, 15, 31, 32, 33, 34]. In the next sections, we examine it in the context of changes in binding affinity.

Note that these imbalances were observed in S2536, but also occur among all single-site mutations of the SKEMPI 2.0 dataset (see Supplementary Section 3).



(a) (b)

Figure 1: Characteristics of the S2536 dataset. (a) Number of occurrences of mutation types; (b) Distribution of the experimental $\Delta\Delta G_b$ values (in kcal/mol).

3.3 Performances on S2536

We tested the performances of the six selected predictors described in Methods (Section 2) on the direct and reverse mutations of the S2536 benchmark dataset. For that purpose, we used the Pearson correlation coefficient between predicted and experimental $\Delta\Delta G_b$ values (ρ) as performance metric. The results are represented in Fig. 3-4. Other metrics such as the root mean squared error and the Spearman rank correlation lead to the same conclusions (as shown in https://github.com/3BioCompBio/DDGb_bias).

This benchmark, though informative, should be considered with caution, as the extent of cross-validation differs according to the predictor. The main issue is that each of the benchmarked predictors is trained on a different subset of S2536, with various covering ratios (CR) with respect to the subset of direct (S2536-D) and reverse (S2536-R) mutations (Table 1). For instance, the training set of mCSM-PPI2 contains 99% of the S2536-D mutations, while that of NetTree only 10%. Furthermore, mCSM-PPI2 is trained on almost all reverse mutations of S2536-R and MutaBind2, on the fraction necessary to balance the number of stabilizing and destabilizing mutations.

The best performing predictors on the direct mutation set S2536-D are mCSM-PPI2, MutaBind2 and SAAMBE-3D with Pearson correlations of 0.91, 0.90 and 0.88, respectively. These values exceed or are very close to the upper bound of 0.89 (Eq. (4)), which indicates overfitting towards the training set. They are followed by SSIPe, BeAtMuSiC and NetTree.

We observe that the performance of all predictors but SSIPe significantly drops when tested on reverse S2536-R mutations. The magnitude of the drop indicates how much each predictor is biased towards direct mutations, which are mostly destabilizing. mCSM-PPI2 and MutaBind2 perform the best on S2536-R, which is expected since they have reverse mutations in their training set; the performance of mCSM-PPI2 drops less than that of MutaBind2, probably because the latter have seen only a part

of the reverse mutations during training (Fig. 3).

Surprisingly, SSIPe is the most robust towards reverse mutations with almost no drop in performance although it does not use reverse mutations in training; its robustness is not acquired by training but rather by the symmetry properties of the model. In contrast, BeAtMuSiC, SAAMBE-3D and NetTree basically fail to predict the $\Delta\Delta G_b$ of reverse mutations. Note the particularly huge drop in performance of SAAMBE-3D, whose Pearson correlation decreases from 0.88 to 0.11; this predictor appears thus to be heavily biased towards destabilizing mutations.

This first benchmark shows that a bias toward destabilizing mutations is present in the context of $\Delta\Delta G_b$ predictions. Note that the drop in performance observed when passing from direct to reverse mutations can partly be attributed to this bias but also to the presence of a larger proportion of mutations in S2536-R than in S2536-D which are unseen during training.

We need to make two important comments at this stage. First, the covering ratio CR between training and benchmark datasets accurately predicts the performances of the predictors. Indeed, we found an almost linear relationship between the CR and the Pearson correlation ρ of the six predictors assessed on the S2536-D set, with a coefficient of determination R^2 as high as 0.81 (Fig. 2). This observation is bothering and raises concerns about the dataset dependency of the predictors and their ability to generalize to mutations outside the training set. This will be further discussed in the next subsection.

The second remark concerns the specificities of BeAtMuSiC. It is the oldest of the tested predictors (Table 1) and the only that was trained on SKEMPI 1.0 mutations, before the SKEMPI 2.0 release. Moreover, it is not completely trained on this mutation set. Indeed, it predicts $\Delta\Delta G_b$ values as a linear combination of changes in folding free energy of the protein complex and of the separate partners (predicted by the PoPMuSiC algorithm [27]), and only uses SKEMPI's $\Delta\Delta G_b$ values to identify the two parameters of the linear combination. BeAtMuSiC's CR is thus not representative of its degree of knowledge of S2536-D mutations. As a matter of fact, the linear relation between the predictors' CR and ρ is even higher if we drop BeAtMuSiC, with a R^2 value of 0.94.

Table 1: Year of publication of the six benchmarked predictors and covering ratio (CR) of their training sets with respect to S2536-D and S2536-R.

Predictor	Year	CR S2536-D	CR S2536-R
mCSM-PPI2	2019	0.99	0.99
MutaBind2	2020	0.75	0.20
BeAtMuSiC	2013	0.62	0.00
SSIPe	2019	0.50	0.00
SAAMBE-3D	2020	0.90[1]	0.00
NetTree	2020	0.10	0.00

[1] Since SAAMBE-3D does not provide its training set, its CR is estimated from the dataset description in its publication [11].

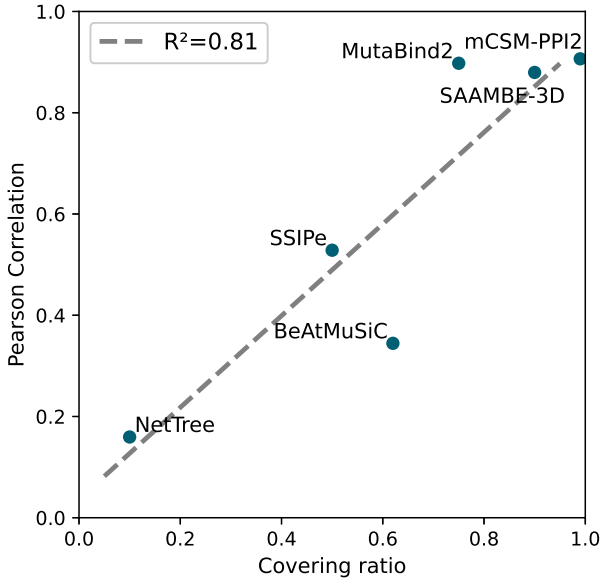


Figure 2: Relation between the covering ratio CR and Pearson correlation ρ between predicted and experimental $\Delta\Delta G_b$ values on the S2536-D set for the six benchmarked predictors. The linear regression line (dashed) and coefficient of determination (R^2) are indicated.

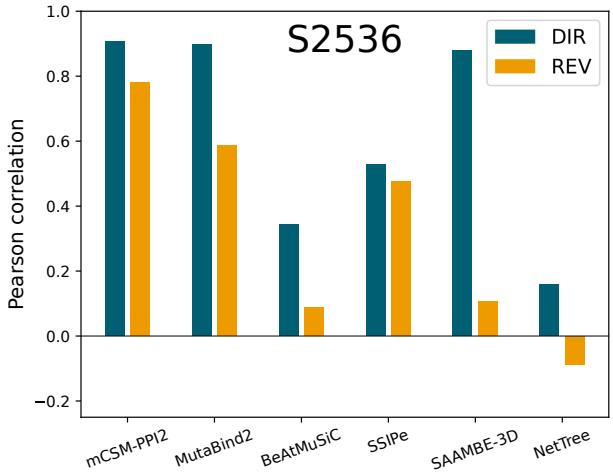


Figure 3: Pearson correlation ρ between the predicted and experimental $\Delta\Delta G_b$ values on C380-D (in blue) and C380-R (in orange), for the six benchmarked predictors.

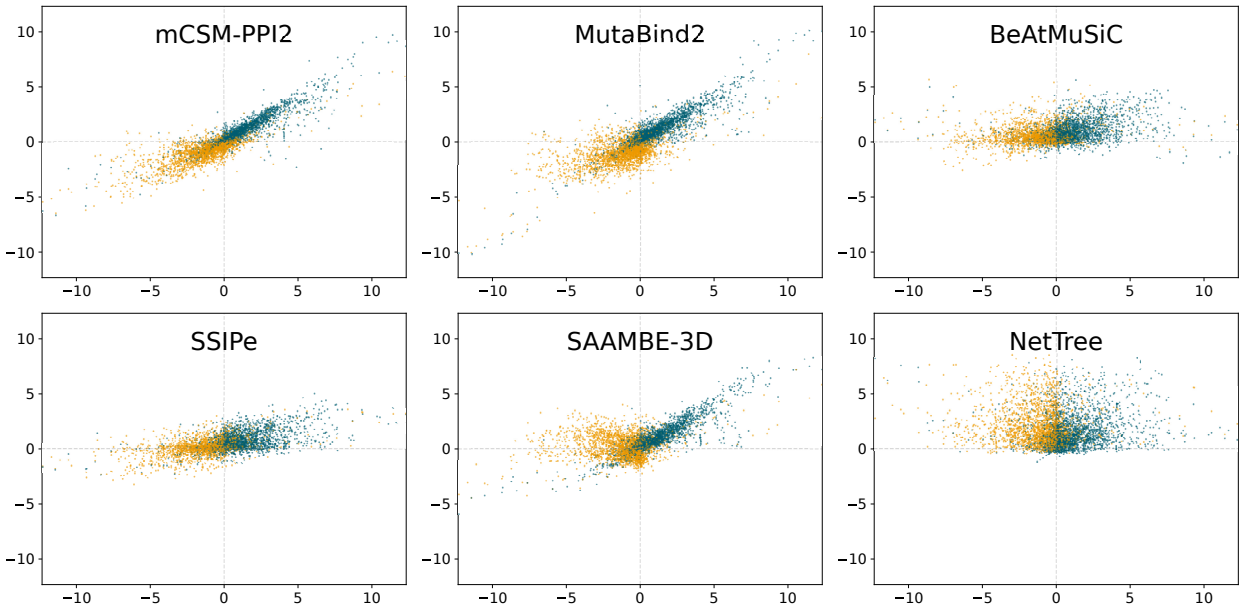


Figure 4: Predicted $\Delta\Delta G_b$ values as a function of experimental $\Delta\Delta G_b$ values (in kcal/mol) for the datasets S2536-D (blue dots) and S2536-R (orange dots). Predictions are obtained with mCSM-PPI2, MutaBind2, BeAtMuSiC, SSIPe, SAAMBE-3D and NetTree.

3.4 Performances on SARS-CoV-2 mutations

The C380 dataset has two major advantages over S2536: it is unknown to the six benchmarked predictors and it is systematic in terms of mutation types. This makes it a better dataset to evaluate the performances of the predictors. By comparing performances on direct and reverse mutations from C380-D and C380-R, we further explored the predictors' bias towards destabilizing mutations; by comparing performances on mutations from S2536 and C380, we estimated the dataset dependence of the predictors.

Since the changes in affinity values reported in C380 were obtained by deep mutagenesis experiments, we cannot assume that they are exactly linearly related with the actual $\Delta\Delta G_b$ values. For instance, values from the whole set of SARS-CoV-2 mutations [18] have a minimal and a maximal experimental bound (see Supplementary Section 3). Because of that, the Pearson correlation and the root mean squared error are inadequate evaluation metrics and we rather chose the Spearman's rank correlation (r). Results of the other metrics on C380 are graphically represented in Supplementary Fig. S-4 and available on https://github.com/3BioCompBio/DDGb_bias.

As shown in Fig. 5, the performances of all predictors drop from S2536 to C380, with no score higher than 0.6 on C380-D and 0.4 on C380-R. This illustrates the heavy impact of the test dataset on the computed accuracy, especially for the best performing predictors mCSM-PPI2 and MutaBind2. They are still the best, with $r = 0.48$ and $r = 0.58$, respectively, on direct mutations, but their lead is considerably reduced. Note that MutaBind2 outperforms mCSM-PPI2 on both direct and reverse C380 mutations, which makes it slightly less dataset dependant and more generalizable. They are followed by SSIPe and BeAtMuSiC, with $r = 0.41$ and $r = 0.34$, respectively, on direct mutations. These two predictors appear to be the most robust against dataset

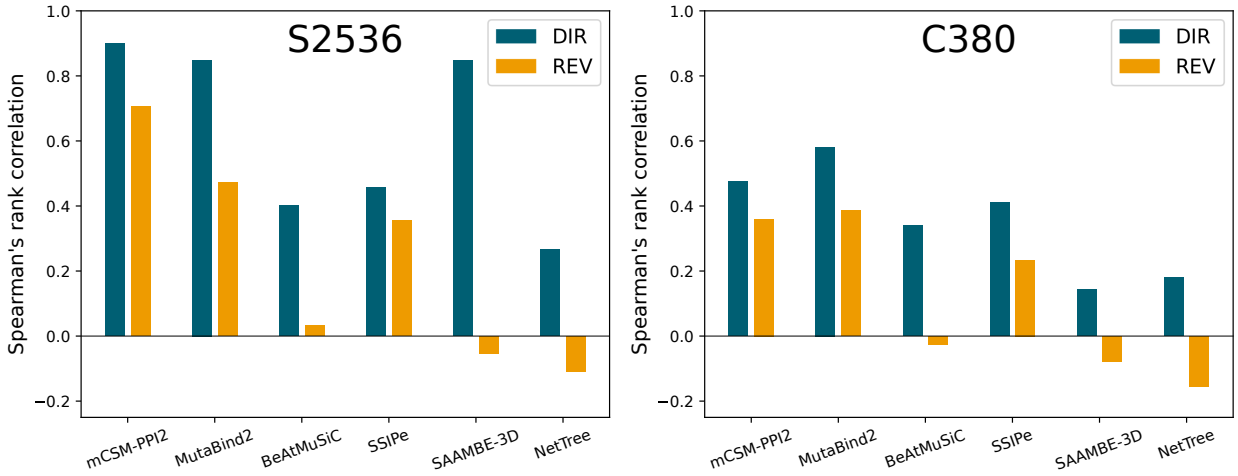


Figure 5: Spearman's rank correlations r between experimental and predicted $\Delta\Delta G_b$ values on direct (in blue) and reverse (in orange) mutations of S2536 and C380.

dependency since their scores remain very similar between S2536-D and C380-D. The lowest prediction scores are those of NetTree and SAAMBE-3D. The latter suffers the most from dataset dependency since it was among the best on S2536-D.

The performance comparison between direct and reverse mutations of C380-D and C380-R confirms the conclusions of the previous section: all predictors suffer, to a different extent, from a bias towards destabilizing mutations. A way to quantify this bias for a given predictor is to compute the symmetry violation defined by Eq. (3) by computing the shift δ defined as:

$$\delta := \Delta\Delta G_b^{wt\to mt} + \Delta\Delta G_b^{mt\to wt}, \tag{6}$$

averaged over all C380 dataset entries. While some fluctuations in δ are expected and acceptable, a systematic deviation of the mean shift $\langle\delta\rangle$ from zero quantifies the asymmetry of a predictor and its bias towards stabilizing or destabilizing mutations. A perfect unbiased value for $\langle\delta\rangle$ is zero; its "worst-case" value can be estimated as twice the average $\Delta\Delta G_b$ value in the dataset of direct mutations, which is 1.24 kcal/mol in C380. We thus estimated the "worst-case" δ-value to be about 2.5 kcal/mol.

We show in Fig. 6 the distributions of δ-values for the six predictors on C380. Analogous δ-values distributions are depicted for S2536 in Supplementary Fig. S-5. We observe that all predictors have a statistically significant shift towards destabilizing mutations, with a vanishing p-value, but amplitude of the shift widely varies. The most symmetric predictors are, as expected, those that perform best on reverse mutations: MutaBind2 with $\langle\delta\rangle = 0.28$ kcal/mol followed by mCSM-PPI2 with $\langle\delta\rangle = 0.47$ kcal/mol.

This confirms that the usage of reverse mutations for training can largely reduce the asymmetry of the predictions. More biased predictions are observed for SSIPe, BeAtMuSiC and SAAMBE-3D, with $\langle\delta\rangle = 1.28$, 1.44 and 1.62 kcal/mol, respectively. These values indicate a bias towards destabilizing mutations, which is, however, still significantly lower than the "worst-case" bias. This means that such predictions are still able to distinguish the tendency between a set of mostly-stabilizing and mostly-destabilizing mutations. In contrast, NetTree obtains $\langle\delta\rangle = 4.05$ kcal/mol, which is

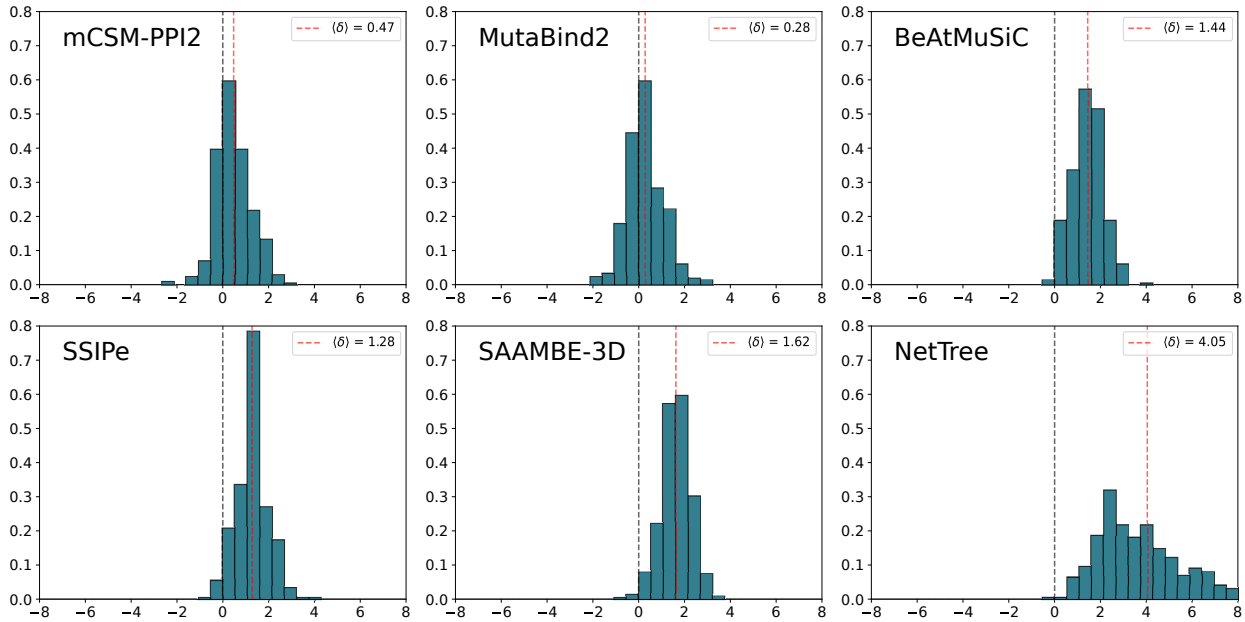


Figure 6: Distribution of the shift δ (in kcal/mol) for the six benchmarked predictors calculated for mutations from C380. The vertical blue dashed lines indicate $\delta = 0$ and the vertical red dashed lines, the value of $\langle\delta\rangle$.

largely above the "worst-case" bias and reflects its inability to distinguish stabilizing from destabilizing mutations. This particularly large $\langle\delta\rangle$-value can partly be explained by NetTree's tendency to predict very large $\Delta\Delta G_b$ values of about 2 kcal/mol, much higher than average experimental values.

In summary, this benchmark represents a fair and objective way to evaluate the performance of the predictors, since C380 is unknown to all. It confirms the presence of biases towards destabilizing mutations in the state-of-the-art $\Delta\Delta G_b$ predictors and highlights the two predictors mCSM-PPI2 and MutaBind2 that are the least affected by this bias. Moreover, our benchmark analysis shows an important dataset dependency of all predictors and demonstrates that high performances on S2536 are overoptimistic and partially due to overfitting. Indeed, the accuracy of some predictors on S2536 even exceeds the theoretical upper bound of the Pearson correlation coefficient ($\sup(\rho) = 0.89$, see Eq. (4)). These high scores obviously do not generalize to mutations outside S2536, for example to C380 mutations, which leaves room for improvement.

3.5 Performances and biases towards mutation properties

We investigated the predictors' performances on subsets of S2536-D containing mutations sharing similar properties, i.e., mutation type, mutation location and type of complex, in order to highlight the predictors' strengths and weaknesses. As the standard deviations σ of the experimental $\Delta\Delta G_b$ values widely differ according to the subset, we used the normalized RMSE defined as nRMSE := RMSE/σ to assess the predictions. The results are shown in Fig. 7. All observations discussed below are statistically significant with almost vanishing p-values (< 0.0001).

We first analyzed separately the subset of mutations towards alanine and the subset

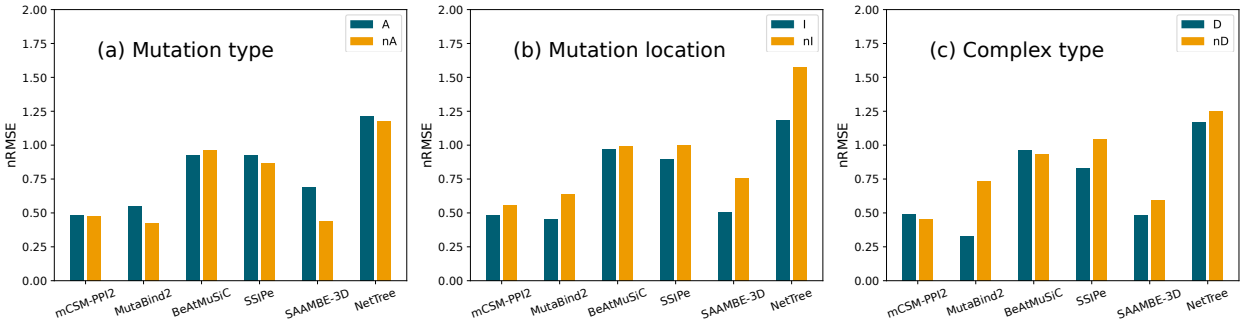


Figure 7: Normalized root mean squared error (nRMSE) of the six predictors on subsets of S2536-D. Subsets were defined based on (a) mutation type: mutation towards Ala (A) vs other mutations (nA); (b) mutation location: mutations at the interface (I) vs other mutations (nI). (c) complex type: mutation on dimeric complexes (D) vs mutations on multi-n-meric complexes ($n > 2$) (nD).

of other mutations. As seen in Fig. 7a, no substantial differences are observed between these two subsets, except that MutaBind2 and SAAMBE-3D perform slightly better on the latter subset. This might be explained by actual strengths/weaknesses of the predictors or could suggest overfitting, since it is easier to memorize $\Delta\Delta G_b$ values on underrepresented mutation types.

Almost all predictors are slightly weaker on mutations outside the protein-protein interface (Figure 7b). This is foreseeable, since effects on binding affinity of non-interface mutations are indirect an thus more difficult to predict. MutaBind2, SAAMBE-3D and NetTree suffer from the largest increase in nRMSE. In contrast, BeAtMuSiC performs the same way on both subsets. SSIPe shows a surprisingly small drop in performance on mutations outside the interface, although it explicitly claims to be only able to predict interface mutations.

When comparing mutations in dimers to mutations in higher-order multimers (Fig. 7c), we observe that mCSM-PPI2 and BeAtMuSiC are the most stable and that MutaBind2, SSIPe and SAAMBE-3D show the largest performance drop. SSIPe's poor performance on higher-order multimers is not surprising as it explicitly announces not to predict such mutations. Paradoxically, mCSM-PPI2 does not require specifying which chains make up the two interactants, although higher-order multimers have several protein-protein interfaces and so there is ambiguity. In spite of this, it maintains the same performance on both subsets, which could suggest overfitting towards its training dataset. In contrast, MutaBind2 asks the chains included in the interactants, but has the largest performance drop on higher-order multimers.

We also assessed the performances on other S2536-D subsets, partitioned by secondary structure, solvent exposure in the complex and interface sub-regions [35] (definitions in Supplementary Section 5), but no relevant observations where found. Results are available at https://github.com/3BioCompBio/DDGb_bias).

3.6 Strategies for avoiding biased predictions

To ensure the generalizability of the predictions, k-fold cross-validation procedures should be carefully performed, avoiding blindly splitting the training set. Indeed,

when separating a dataset into folds, a direct mutation and its corresponding reverse mutation should end up in the same fold to avoid that information from one mutation influences the prediction of the other. As the S2536 dataset contains multiple homologous complexes differing by only a few mutations, random cross-validations can also lead to information leaks from training to testing sets and provide overoptimistic results. Thus, mutations on homologous complexes should also be kept in the same fold [13].

However, dataset biases can be learned by the predictors even if a strict cross-validation procedure is used. To illustrate this, we started by noticing that half of the mutations from S2536-D are towards alanine ($X \rightarrow A$) and thus that half of the mutations from S2536-R are from alanine ($A \rightarrow X$). Knowing moreover that S2536-D and S2536-R contain mostly destabilizing and mostly stabilizing mutations, respectively, the sign of $\Delta\Delta G_b$ can be often correctly guessed for $X \rightarrow A$ and $A \rightarrow X$ mutations while holding no predictive power. In other words, predictors can learn imbalances and cross correlations between mutations' properties from S2536, which improves its performances in cross-validation while also increasing its generalization error.

As a proof of this phenomenon, we created a "perfectly biased" predictor, which estimates $\Delta\Delta G_b$ as the mean of the experimental $\Delta\Delta G_b$ values of the same mutation type in the training set (or zero if the mutation type was never encountered). This predictor manages to obtain a Pearson correlation $\rho = 0.46$ on S2536-DR in 10-fold cross-validation. When applying the same predictor (trained on S2536-DR) on mutation type-balanced, interface-only entries from C380-DR, the Pearson correlation falls to $\rho = 0.35$, and completely vanishes when dropping the interface filter and applying the predictor to the whole dataset of mutations on the RBD-ACE2 complex (-DR) with $\rho = 0.04$. The same phenomenon also happens, with however slightly smaller correlations, when considering direct mutations only. We indeed found $\rho = 0.34$ in 10-fold cross-validation on S2536-D, $\rho = 0.27$ on C380-D and $\rho = 0.05$ on RBD-ACE2 (-D). Note that these scores are only an underestimation of how dataset-dependent cross correlations from S2536 can impact predictions; we have indeed only considered mutation type-related biases.

As extensively discussed above, asymmetric predictions are another type of unwanted bias. One easy way to avoid it is to symmetrize the prediction results. Indeed, the prediction shift δ vanishes when redefining the prediction of a mutation $wt \rightarrow mt$ as:

$$\overline{\Delta\Delta G_b}^{wt \rightarrow mt} := \frac{\Delta\Delta G_b^{wt \rightarrow mt} - \Delta\Delta G_b^{mt \rightarrow wt}}{2}, \tag{7}$$

with, as a consequence, $\delta = \overline{\Delta\Delta G_b}^{wt \rightarrow mt} + \overline{\Delta\Delta G_b}^{mt \rightarrow wt} = 0$. This operation requires both wild-type and mutant structures, but does not introduce any internal modifications to the predictor itself. Some but not all mutant structures have been resolved experimentally; we listed in the https://github.com/3BioCompBio/DDGb_bias repository the pairs of resolved wild-type and mutant structures from SKEMPI 2.0 that are separated by a single mutation (more details in Supplementary Section 6). Alternatively, the unavailable mutant structures can be modeled with homology modeling techniques using the wild-type structure as a template.

Symmetrized versions of all tested predictors were obtained using Eq. (7). For predictors that suffer from a strong bias towards destabilizing mutations, the Pearson

correlation coefficient of the symmetrized version falls somewhere between their scores on direct and on reverse mutations. In contrast, the three most symmetric predictors, mCSM-PPI2, MutaBind2 and SSIPe, show an improved score on the reverse datasets S2536-R and C380-R, as well on the combined datasets S2536-DR and C380-DR, and similar or only slightly lower performance on the direct datasets S2536-D and C380-D (Supplementary Section 4). This shows that the overall performances of some predictors can be improved while also increasing their symmetry without introducing any internal changes to the model.

As seen in the previous subsections, an alternative strategy to reduce the asymmetry of the predictions consists in using reverse mutations for training. Among the tested predictors, MutaBind2 and mCSM-PPI2 apply this technique and reach good symmetry properties. This practice increases the generalizability and robustness of predictors. However, the symmetrization of the training set has to be done carefully. Indeed, due to the presence of wild-type/mutant pairs in SKEMPI 2.0, adding the reverse of all mutations, as done in mCSM-PPI2, leads to redundant entries that should be avoided, as they are a source of biases.

3.7 Predictors' computational efficiency

Computational time efficiency is another characteristic to consider when choosing a prediction method, especially when a large set of mutations has to be analyzed, as for example in the study of variants impact on the interactome [2]. In terms of speed, BeAtMuSiC and SAAMBE-3D are the only of the tested predictors that are fast enough to enable large-scale computational mutagenesis experiments; indeed, they are able to predict all possible single-site mutations in a protein complex in a few to a few tens of seconds. In contrast, mCSM-PPI2, MutaBind2, SSIPe and NetTree are time consuming and require tens of seconds to tens of minutes just to run a single mutation. This prevents their use for large-scale applications.

4 Conclusions

Our benchmarks clearly illustrate that the predictive power of a method is not necessarily well represented by its scores on its training dataset even if a strict cross-validation procedure is used. This makes the validation process particularly challenging. Here we identified two main issues, which are the predictors' systematic asymmetry and their lack of generalization on mutations outside their training set. They are discussed below.

Lack of generalization. A major challenge in $\Delta\Delta G_b$ predictions is to distinguish between statistical relations that are dataset dependent and the "true" ones that have a biological meaning. We would like to stress that, while physics- and evolution-based methods are at least partly equipped to tackle this problem, pure machine learning methods struggle to make this distinction. This can explain the particularly large performance drop on unknown mutations observed for SAAMBE-3D and NetTree, which are the two purest machine learning predictors. The generalizability of a predictor must be tested on independent sets of mutations outside the training set. Sets of systematic mutations obtained by deep mutagenesis experiments, such as C380, have the advantage of not being impacted by literature biases. They are thus appropriate

for validating and benchmarking predictions, even though their $\Delta\Delta G_b$ values are less accurate than those obtained by individual thermodynamic experiments.

Symmetry properties. Symmetry properties should be carefully checked when constructing a prediction model. One way to assess them is on the basis of the shift δ (Eq. (6)). As a general rule, the symmetry of a predictor can be achieved by: (1) using symmetric data during training by including all or a fraction of reverse mutations, as done in mCSM-PPI2 and MutaBind2; (2) enforcing symmetry in the predictor's mathematical model, as in [15]; and (3) applying symmetry-correction methods through, e.g., the symmetrization defined in Eq. (7). Method (1) is a good practice which, as we showed, can increase the generalizability of the predictions. Method (2) can help the predictor to be symmetric, but it is only applicable when the mathematical expression of the model is known. Method (3) is the easiest to implement, but is efficient only if the predictor is already robust to symmetry.

There are additional challenges that need to be addressed. First, further data on binding affinity and interactions need to be collected. Accurate $\Delta\Delta G_b$ thermodynamics data have not been systematically collected for the past five years, after SKEMPI 2.0's release. Also, deep mutagenesis data of binding affinity are currently generated at a high rate but need to be collected, curated and harmonized. Secondly, interpretation of $\Delta\Delta G_b$ predictions is an issue that we do not explore in this paper and that is not sufficiently discussed in the literature.

Finally, there is a need for more independent assessments. We invite the community to set up blind challenges for the prediction of changes in protein-protein binding affinity upon mutations, similar to what has been done during the 26th critical assessment of predicted interactions (CAPRI) experiment [36]. These community-wide blind challenges provide important insights into whether and how different predictors achieve the targeted accuracy, and help drive the development of new methods.

References

[1] Nidhi Sahni, Song Yi, Mikko Taipale, Juan I Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I Karras, Yang Wang, et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660, 2015.

[2] Feixiong Cheng, Junfei Zhao, Yang Wang, Weiqiang Lu, Zehui Liu, Yadi Zhou, William R Martin, Ruisheng Wang, Jin Huang, Tong Hao, et al. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature genetics*, 53(3):342–353, 2021.

[3] Laura Nevola and Ernest Giralt. Modulating protein–protein interactions: the potential of peptides. *Chemical Communications*, 51(16):3302–3315, 2015.

[4] Haiying Lu, Qiaodan Zhou, Jun He, Zhongliang Jiang, Cheng Peng, Rongsheng Tong, and Jianyou Shi. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal transduction and targeted therapy*, 5(1):1–23, 2020.

[5] Stephanie Leavitt and Ernesto Freire. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current opinion in structural biology*, 11(5):560–566, 2001.

[6] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[7] Carlos HM Rodrigues, Yoochan Myung, Douglas EV Pires, and David B Ascher. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47(W1):W338–W344, 2019.

[8] Ning Zhang, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R Panchenko, and Minghui Li. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience*, 23(3):100939, 2020.

[9] Yves Dehouck, Jean Marc Kwasigroch, Marianne Rooman, and Dimitri Gilis. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic acids research*, 41(W1):W333–W339, 2013.

[10] Xiaoqiang Huang, Wei Zheng, Robin Pearce, and Yang Zhang. SSIPe: accurately estimating protein–protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. *Bioinformatics*, 36(8):2429–2437, 2020.

[11] Swagata Pahari, Gen Li, Adithya Krishna Murthy, Siqi Liang, Robert Fragoza, Haiyuan Yu, and Emil Alexov. SAAMBE-3D: predicting effect of mutations on protein–protein interactions. *International journal of molecular sciences*, 21(7):2563, 2020.

[12] Menglun Wang, Zixuan Cang, and Guo-Wei Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2):116–123, 2020.

[13] Iain H Moal and Juan Fernández-Recio. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, 28(20):2600–2607, 2012.

[14] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

[15] Fabrizio Pucci, Katrien Bernaerts, Fabian Teheux, Dimitri Gilis, and Marianne Rooman. Symmetry principles in optimization problems: an application to protein stability prediction. *IFAC-PapersOnLine*, 48(1):458–463, 2015.

[16] Dinara R Usmanova, Natalya S Bogatyreva, Joan Ariño Bernad, Aleksandra A Eremina, Anastasiya A Gorshkova, German M Kanevskiy, Lyubov R Lonishin, Alexander V Meister, Alisa G Yakupova, Fyodor A Kondrashov, et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics*, 34(21):3653–3658, 2018.

[17] Fabrizio Pucci, Katrien V Bernaerts, Jean Marc Kwasigroch, and Marianne Rooman. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*, 34(21):3659–3665, 2018.

[18] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310, 2020.

[19] George D Rose, Ari R Geselowitz, Glenn J Lesser, Richard H Lee, and Micheal H Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, 1985.

[20] Georgios A Dalkas, Fabian Teheux, Jean Marc Kwasigroch, and Marianne Rooman. Cation–π, amino–π, π–π, and H-bond interactions stabilize antigen–antibody interfaces. *Proteins: Structure, Function, and Bioinformatics*, 82(9):1734–1746, 2014.

[21] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[22] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[23] Stephen M Lu, Wuyuan Lu, MA Qasim, Stephen Anderson, Izydor Apostol, Wojciech Ardelt, Theresa Bigler, Yi Wen Chiang, James Cook, Michael NG James, et al. Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *Proceedings of the National Academy of Sciences*, 98(4):1410–1415, 2001.

[24] Henrik Gardsvoll, Bernard Gilquin, Marie Hélène Le Du, Andre Ménèz, Thomas JD Jørgensen, and Michael Ploug. Characterization of the functional epitope on the urokinase receptor: complete alanine scanning mutagenesis supplemented by chemical cross-linking. *Journal of Biological Chemistry*, 281(28):19260–19272, 2006.

[25] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807):215–220, 2020.

[26] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using MODELLER. *Current protocols in bioinformatics*, 54(1):5–6, 2016.

[27] Yves Dehouck, Aline Grosfils, Benjamin Folch, Dimitri Gilis, Philippe Bogaerts, and Marianne Rooman. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, 25(19):2537–2543, 2009.

[28] Sarah Sirin, James R Apgar, Eric M Bennett, and Amy E Keating. AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[29] Ludovica Montanucci, Pier Luigi Martelli, Nir Ben-Tal, and Piero Fariselli. A natural upper bound to the accuracy of predicting protein stability changes upon mutations. *Bioinformatics*, 35(9):1513–1517, 2019.

[30] Silvia Benevenuta and Piero Fariselli. On the upper bounds of the real-valued predictions. *Bioinformatics and Biology Insights*, 13:1177932219871263, 2019.

[31] Octav Caldararu, Rukmankesh Mehra, Tom L Blundell, and Kasper P Kepp. Systematic investigation of the data set dependency of protein stability predictors. *Journal of Chemical Information and Modeling*, 60(10):4772–4784, 2020.

[32] Kristoffer T Bæk and Kasper P Kepp. Data set and fitting dependencies when estimating protein mutant stability: Toward simple, balanced, and interpretable models. *Journal of Computational Chemistry*, 43(8):504–518, 2022.

[33] Vladimir Potapov, Mati Cohen, and Gideon Schreiber. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein engineering, design & selection*, 22(9):553–560, 2009.

[34] Dinara R Usmanova, Natalya S Bogatyreva, Joan Ariño Bernad, Aleksandra A Eremina, Anastasiya A Gorshkova, German M Kanevskiy, Lyubov R Lonishin, Alexander V Meister, Alisa G Yakupova, Fyodor A Kondrashov, et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics*, 34(21):3653–3658, 2018.

[35] Emmanuel D Levy. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology*, 403(4):660–670, 2010.

[36] Rocco Moretti, Sarel J Fleishman, Rudi Agius, Mieczyslaw Torchala, Paul A Bates, Panagiotis L Kastritis, Joao PGLM Rodrigues, Mikaël Trellet, Alexandre MJJ Bonvin, Meng Cui, et al. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 81(11):1980–1987, 2013.

[37] Sanghee Yoo, David G Myszka, Chin-yah Yeh, Maureen McMurray, Christopher P Hill, and Wesley I Sundquist. Molecular recognition in the HIV-1 capsid/cyclophilin A complex. *Journal of molecular biology*, 269(5):780–795, 1997.

[38] Theresa R Gamble, Felix F Vajdos, Sanghee Yoo, David K Worthylake, Megan Houseweart, Wesley I Sundquist, and Christopher P Hill. Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell*, 87(7):1285–1294, 1996.

[39] Minghui Li, Franco L Simonetti, Alexander Goncearenco, and Anna R Panchenko. Mutabind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic acids research*, 44(W1):W494–W501, 2016.

[40] Kurt S Thorn and Andrew A Bogan. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, 2001.

[41] Song Liu, Chi Zhang, Hongyi Zhou, and Yaoqi Zhou. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics*, 56(1):93–101, 2004.

[42] Panagiotis L Kastritis, Iain H Moal, Howook Hwang, Zhiping Weng, Paul A Bates, Alexandre MJJ Bonvin, and Joël Janin. A structure-based benchmark for protein–protein binding affinity. *Protein Science*, 20(3):482–491, 2011.

[43] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[44] Sherlyn Jemimah, K Yugandhar, and M Michael Gromiha. PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, 33(17):2787–2788, 2017.

[45] Douglas EV Pires, David B Ascher, and Tom L Blundell. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342, 2014.

[46] Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A Bates, Juan Fernandez-Recio, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, 427(19):3031–3041, 2015.

[47] Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. BindProfX: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology*, 429(3):426–434, 2017.

[48] MD Shaji Kumar and M Michael Gromiha. PINT: protein–protein interactions thermodynamic database. *Nucleic acids research*, 34(suppl_1):D195–D198, 2006.

[49] Quanya Liu, Peng Chen, Bing Wang, Jun Zhang, and Jinyan Li. dbmpikt: a database of kinetic and thermodynamic mutant protein interactions. *BMC bioinformatics*, 19(1):1–7, 2018.

[50] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002.