This version of the manuscript corresponds to the pre-print version of the following paper:

C. Geng, A. Vangone and A.M.J.J. Bonvin. 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*, *29*, 291-299 (2016).

http://dx.doi.org/10.1093/protein/gzw020

Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes.

Cunliang Geng, Anna Vangone and Alexandre M.J.J. Bonvin*

Computational Structural Biology Group, Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands.

* Phone: +31.30.2533859, Fax: +31.30.2537623, Email: a.m.j.j.bonvin@uu.nl

Running title (<50 characters):
Impact of experimental methods on ΔΔG predictors

Abstract

Reliable prediction of binding affinity changes (ΔΔG) upon mutations in protein complexes relies not only on the performance of computational methods but also on the availability and quality of experimental data. Binding affinity changes can be measured by various experimental methods with different accuracies and limitations. To understand the impact of these on the prediction of binding affinity change, we present the Database of binding Affinity Change Upon Mutation (DACUM), a database of 1872 binding affinity changes upon single point mutations, a subset of the SKEMPI database (Moal and Fernández-Recio, 2012) extended with information on the experimental methods used for ΔΔG measurements. The ΔΔG data were classified into different datasets based on the experimental method used and the position of the mutation (interface and non-interface). We tested the prediction performance of the original HADDOCK score, a newly trained version of it and mCSM (Pires et al., 2014), one of the best reported ΔΔG predictor so far, on these various datasets. Our results demonstrate a strong impact of the experimental methods on the performance of binding affinity change predictors for protein complexes. This underscores the importance of properly considering and carefully choosing experimental methods in the development of novel binding affinity change predictors. The DACUM database is available online at https://github.com/haddocking/DACUM.

Keywords: binding affinity/computational prediction/experimental methods/protein-protein interactions/singe point mutation

1. Introduction

Protein-protein interactions (PPIs) play fundamental roles in the regulation of various biological processes. A single mutation in a protein can be sufficient to alter the properties of its PPIs network by modulating the strength of interaction with its partners, as measured by the binding affinity, or binding affinity changes upon mutation, which, in turn, can lead to malfunction or disease (Stites, 1997). Studying and measuring the impact that mutations might have on binding affinity is therefore crucial to a proper understanding of PPIs and related biological phenomena.

Currently, there are many experimental methods that can be used to measure the binding affinity of a protein complex, for example, isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), fluorescence, spectroscopy and stopped-flow assays (Kastritis and Bonvin, 2013). These methods have been extensively applied to study the difference of binding affinity between wild type and mutant protein complexes, expressed as binding affinity changes upon mutation ($\Delta\Delta G$). During the last decades, a large amount of experimental data on binding affinity change upon mutation have been generated. The most recent and largest collection of such data is the SKEMPI database (Structural database of Kinetics and Energetics of Mutant Protein Interactions) (Moal and Fernández-Recio, 2012), which contains 2317 entries with available thermodynamics parameters for both wild type (native) and mutant protein-protein complexes.

As experimentally measuring the effect of mutations on protein complexes can be costly and time-consuming, scientists have also turned to computational methods to predict binding affinity changes upon mutation in protein complexes. Taking advantage of the SKEMPI database, several binding affinity change predictors have been developed over the last years (Dehouck et al., 2013; Moal and Fernández-Recio, 2013; Berliner et al., 2014; Dourado and Flores, 2014; Li et al., 2014; Pires et al., 2014; Brender and Zhang, 2015). Of these, mCSM (mutation Cutoff Scanning Matrix) (Pires et al., 2014) shows a strong performance with an r of 0.81 in terms of Pearson's correlation coefficient between the predicted and the measured binding affinity changes. It was trained on a dataset of single point mutations derived from SKEMPI. Although datasets collecting mutation data are essential to develop and test new prediction approaches, one should also pay attention to the quality of the data. By classifying the latter based upon the experimental methods used for their measurements, we have recently

demonstrated that these have an impact on the achievable accuracy of predictor of absolute binding affinities (Vangone and Bonvin, 2015). In the case of binding affinity changes upon mutation, none of this has yet been investigated, mainly because SKEMPI does not provide information on the experimental methods used in their measurements.

Besides the experimental method used for its measurement, also the position of a mutation may have an impact on the achievable accuracy of a predictor. According to Levy's classification of structural regions (Levy, 2010), a mutation located in the interface core, rim or support regions can be defined as interface mutation, while one on the remaining surface or in the interior region (buried mutation) is considered a non-interface mutation. It is easy to assume that an interface mutation will have a larger and more direct effect on PPIs than a non-interface one, therefore causing a larger change in binding affinity. It seems therefore necessary, when training and testing binding affinity change predictors, to also distinguish the data based on their location. Despite the fact that this is provided in SKEMPI, most known SKEMPI-based predictors have not taken that information into account.

Several approaches have been proposed to predict from structure the impact of mutations on binding affinity. Several of those implement empirical potentials (Li et al., 2014; Brender and Zhang, 2015). They are typically composed of physico-chemical terms such as van der Waals, electrostatic and desolvation energies in combination with various other terms such as changes in buried surface area. Rather similar functions are used in our docking software HADDOCK (Dominguez et al., 2003) to rank docking models. Despite its simplicity the HADDOCK scoring function has proven robust for scoring as demonstrated in the joint CASP-CAPRI prediction round (Lensink et al., 2016).

Here we present DACUM, a Database of binding Affinity Change Upon Mutation derived from SKEMPI, which, in addition to the information reported in SKEMPI, also provide the information on the experimental methods used to measure the binding affinity for each mutation, and a classification of the mutations as interface/non-interface (which effectively represents as simplification of the SKEMPI classification from rim, core and support into interface, and interior and surface regions into non-interface). We test the impact of the experimental methods used for affinity measurement and of the location of the mutations on the achievable accuracy of predictors of binding affinity changes. This is done using mCSM, one of the best performing predictor to date and our simple HADDOCK score. Considering

that the scoring function of HADDOCK has been initially conceived to score and rank docking poses, we retrained it specifically to predict the impact of mutations. Our overall results demonstrate that experimental methods strongly impact the ΔΔG prediction performance, as was previously observed for absolute affinity prediction (Vangone and Bonvin, 2015). As a consequence, experimental data sets should be carefully chosen when training new predictors.

2. Materials and Methods

Building the database. Our Database of binding Affinity Change Upon Mutation (DACUM) was built from the SKEMPI database (Moal and Fernández-Recio, 2012). SKEMPI contains data on thermodynamic and kinetic parameters for protein-protein interactions for which the structure of the complex is available in the Protein Data Bank (Berman et al., 2000). Version 1.1 contains 3047 mutations for 158 complexes (PDB IDs) including single and multiple point mutations, with associated dissociation constant (K_d) together with the original reference (PubMed ID) and the location of the mutated residue, according to the definition provided in Levy's method (Levy, 2010).

In DACUM we added data about the experimental method used to measure binding affinity changes for each single point mutant. This was done by checking manually the original references for the reported mutants. We classified the experimental methods as defined in Table 1. In this process we found several errors in the reported binding affinity values that were corrected in DACUM. These are reported in SI Table S1.

We further filtered and classified the mutations as follows:

1. Only single point mutations were kept.
2. Only mutations with a unique experimental binding affinity value were kept.
3. Mutations were classified according to their location into interface (ITF) and non-interface (NIF) mutations. ITF combines the core (COR), support (SUP) and rim (RIM) classes in SKEMPI, while NIF combines interior (INT) and surface (SUR).
4. ITF and NIF mutations were further sub-divided into different classes based on the experimental methods reported in the related references (see Table 1).

Finally, the changes in binding affinity upon mutation ΔΔG were calculated from the reported K_d and temperature (T) as:

$$\mathbf{\Delta G = RT \ln(K_d/C^0)} \quad \mathbf{(1)}$$

$$\mathbf{\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}} \quad \mathbf{(2)}$$

where the ΔG_{mut} and ΔG_{wt} are the binding free energies of the mutant and wild-type complexes, respectively, *R* the ideal gas constant (0.0019872 kcal K^{-1} mol^{-1}), C^0 the standard reference concentration (1 mol L^{-1}), *T* the temperature in Kelvin, and K_d the dissociation constant in mol L^{-1}.

Our database can be freely downloaded from https://github.com/haddocking/DACUM.

HADDOCK refinement of protein-protein complexes. Starting from the wild-type entries reported in DACUM, we generated models of all reported mutants as follows.

From the reported wild-type PDB structures of the complexes we generated all mutants using our in-house script mutate.py (available from https://github.com/haddocking/haddock-tools), which simply changes the residue name on the PDB entry. The histidine protonation states were defined using molprobity (Davis et al., 2007), the missing atoms were built and the resulting models subjected to a gentle refinement in explicit water (TIP3P) with the OPLS force field (Jorgensen and Tirado-Rives, 1988) using an 8.5Å cut-off, following the default protocol implemented in the refinement interface of the HADDOCK2.2 web server (van Zundert et al., 2016). For each mutant, 50 models were generated in this way. They were ranked using the default HADDOCK Score (HS):

$$\mathbf{HS = 1.0\ E_{vdw} + 0.2\ E_{elec} + 1.0\ E_{desolv}} \quad \mathbf{(3)}$$

where E_{vdw} is the intermolecular van der Waals energy described by a 12-6 Lennard-Jones potential, E_{elec} the intermolecular electrostatic energy described by a Coulomb potential and E_{desolv} an empirical desolvation energy term (Fernández-Recio et al., 2004). Besides these three energy terms, the buried surface area [$Å^2$] (BSA) was also calculated.

For the best HADDOCK model (i.e the one with the lowest HS) of each complex, the HADDOCK score and its components (E_{vdw} , E_{elec}, E_{desolv} and BSA) were collected and differences between mutant and wild-type complexes were calculated as:

$$\Delta HS = HS_{mut} - HS_{wt} \quad \textbf{(4)}$$

$$\Delta E_{vdw} = E_{vdw\text{-}mut} - E_{vdw\text{-}wt} \quad \textbf{(5)}$$

$$\Delta E_{elec} = E_{elec\text{-}mut} - E_{elec\text{-}wt} \quad \textbf{(6)}$$

$$\Delta E_{elec} = E_{elec\text{-}mut} - E_{elec\text{-}wt} \quad \textbf{(7)}$$

$$\Delta BSA = BSA_{mut} - BSA_{wt} \quad \textbf{(8)}$$

These were used for training of a new binding affinity change predictor (see below).

Prediction of affinity changes upon mutation using mCSM. The mCSM webserver (Pires et al., 2014) was used to predict the binding affinity change upon mutation for all 1872 DACUM entries using default settings.

Correlation and regression analysis. The correspondence between experimental and predicted binding affinity changes was measured using Pearson's correlation coefficients (r).

Trained HADDOCK ΔΔG predictor. Based on the various HADDOCK terms, we trained a predictor using multiple linear regression, applying the Akaike information criterion (AIC) in a stepwise selection approach (backward and forward) to identify significant parameters and calculate weights only for them.

$$\mathbf{HS_{\Delta\Delta G} = w_{vdw}\, E_{vdw} + w_{elec}\, E_{elec} + w_{desolv}\, E_{desolv} + w_{BSA}\, BSA + C} \quad \textbf{(9)}$$

Both regression analysis and AIC were implemented in R (R Development Core Team, 2014). The standard error of regression (σ) was calculated to quantify the difference between predicted and experimental values.

3. Results

Database of binding affinity change upon mutation

From the original 3047 mutations reported in SKEMPI, after filtering (see Material and Methods) 1872 single point, non-redundant mutations were selected for DACUM. For each entry in DACUM, the SKEMPI_ID, ΔΔG value, position of mutation, experimental method and related reference are provided. Compared to other datasets (Dehouck et al., 2013; Moal and Fernández-Recio, 2013; Berliner et al., 2014; Dourado and Flores, 2014; Li et al., 2014; Pires et al., 2014; Brender and Zhang, 2015) derived from SKEMPI, DACUM is the first to provide detailed information about experimental methods used to measure the binding affinity for each mutant. The experimental methods reported in DACUM are summarized in Table 1. In total, 15 different experimental methods were obtained from 119 references. This information was used to classify the data into different datasets based on the experimental measurement method. Further, the ΔΔG values in DACUM were calculated and obtained after correcting a few (44) misreported values in SKEMPI (see SI Table S1).

Statistics of methods and interface / non-interface mutations

DACUM contains 1580 and 292 “interface” and “non-interface” mutations, respectively. These two groups (referred to as “ALL” datasets) were further classified according to the experimental binding affinity measurement methods. The resulting ΔΔG distributions and number of mutants for each sub-set are shown in Figure 1.

For interface mutations (Figure 1A), the ALL dataset presents a large distribution of experimental ΔΔG values, ranging from -13 to 13 kcal/mol for 1580 mutations. The number of mutations largely varies from dataset to dataset, e.g. the Spectroscopy (SP) dataset is the largest with 564 mutations, whereas both ELFA and EMSA datasets contain only 6 mutations each. Despite the various distributions for these datasets, most show a range roughly between -4 and 6 kcal/mol with a median of about 1 kcal/mol.

Compared to interface mutations, the various methods datasets for non-interface mutations show much tighter distributions of experimental ΔΔG. Not surprisingly, the variability of ΔΔG is roughly 5 times smaller, ranging from -1 to 1 kcal/mol. Also the sizes of the datasets

are much smaller for a total of 292 non-interface compared to 1580 interface mutations. Further some methods are not represented (SPR, SFSP, IAGE and EMSA).

Other features of the dataset such as number of mutation types, number of complexes and number of mutations per complex (PDB IDs) are summarized in SI Table S2. Most datasets (11/15 and 8/11 for interface and non-interface mutations, respectively) have a ratio of number of mutations per complex smaller than 10, indicating a well-balanced dataset in terms of number of complexes and mutations (i.e. not dominated by a single complex).

ΔΔG prediction performance of mCSM on different method datasets

We tested the performance of mCSM, which was trained on all single point mutations of SKEMPI database, on the various methods datasets. mCSM achieved a correlation of r=0.61 and 0.17 for the ALL datasets of interface mutations and non-interface mutations, respectively (Figure 2A and 2B). The original paper (Pires et al., 2014) reports a correlation of r=0.80 for the 2317 single point mutations dataset in SKEMPI.

When considering the various methods datasets, mCSM showed a similar performance on both interface and non-interface mutations with average correlations of r=0.50±0.25 and 0.42±0.41, respectively. It achieves a correlation higher than 0.5 for 10 out of the 15 interface datasets, and a negative correlation for 1 dataset. The performance varies largely depending on the experimental measurement method ranging from r = -0.140 for ELISA to 0.844 for FL. For non-interface mutations, excluding the IAFL dataset with only 2 mutations, the correlation coefficients exceed 0.5 for 5 datasets with the largest values (r=0.70) for FL, and a negative correlation of r=-0.620 for IASP. These large variations in prediction performance clearly indicate a rather strong impact of the experimental measurement method on the reliability of the prediction.

ΔΔG prediction performance of the raw HADDOCK score on different method datasets

Not surprisingly, the raw HADDOCK score performs rather poorly when it comes to predicting changes in binding affinity upon mutation with a correlation coefficient of r=0.28 for the ALL interface mutations dataset (Figure 2A – green bars). This was already observed for absolute binding affinity prediction of protein-protein complexes (Kastritis and Bonvin,

2010; Kastritis et al., 2011). Only for 4 method datasets we obtained correlation coefficients over 0.5, namely SFFL, SFFP, IAGE and RA. Three method datasets have negative correlations, namely CSPRIA, ELFA and EMSA dataset. The largest correlation was achieved on SFSP dataset with only 10 mutations (r=0.918), while the minimum occurred for ITC dataset with 42 mutations (r=0.076).

The non-interface mutations showed insignificant correlation of r=0.01 for the ALL dataset, with an average value of r=0.08±0.35 for all 11 method datasets (SI Figure S1 – green bars). This is not surprising since the HADDOCK score is only calculated from the intermolecular energies with an 8.5Å cutoff. As such any remote mutations will not really affect the score.

Performance of the trained HADDOCK ΔΔG predictor on different method datasets

Using the components of the HADDOCK score (Evdw, Eelec, Edesolv and BSA) we trained a new ΔΔG predictor. The entire interface dataset and each method dataset were separately used to train a multiple linear regression model (see Materials and Methods). The results of the trained predictors are shown in Table 2. We can see that the HADDOCK terms selected and their corresponding weights for predictors of different method datasets were obviously different, and the training failed to generate a predictor for CSPRIA, ELISA and EMSA datasets with 23, 16 and 6 mutations, respectively. It is noticeable that none of the trained predictors reported all the four HADDOCK terms, and only three or less HADDOCK terms were kept in predictors after applying the feature selection method AIC.

The results for each predictor are shown in Figure 2A (red bars). The predictor trained and tested on the ALL dataset achieved a correlation of r=0.27 with standard error of regression σ=2.03 kcal/mol, which is roughly equivalent to the performance of the raw HADDOCK Score on the ALL dataset. Compared with the ALL predictor, SPR and SP predictors performed similarly on their individual dataset, while other predictors performed much better, e.g. the Pearson's r of ITC, FL and SFFL predictors were about 0.5, and that of SFSP, IAFL, IASP, IAGE, IARA, RA and ELFA predictors were higher than 0.6. The predictor having the largest correlation is SFSP with r=0.939 for a limited set of 10 mutations, while the SPR predictor got the lowest correlation with r=0.234 for 305 mutations. From Figure 2 it might look as if the trained predictor is now performing as well or better than mCSM, but this is misleading since the HADDOCK predictors were trained separately for each dataset while the

same mCSM model was used for all datasets. A comparison of performance of mCSM and the ALL trained HADDOCK predictor is shown in SI Figure S2.

4. Discussion

In this work we reported DACUM, a filtered subset of SKEMPI with 1872 cleaned, single point, non-redundant mutations with additional information on the experimental methods used to measure the binding affinity. This allowed us to define 15 different classes of experimental methods.

The datasets of interface mutations show a larger distribution of binding affinity change than those of non-interface mutations. This is not surprising, since it is expected that mutations at the interface will have a stronger impact on the affinity, resulting in larger changes in binding affinity between wild-type and mutant complexes compared to non-interface mutations. As shown previously, the latter can still contribute to overall binding affinity of a complex (Kastritis and Bonvin, 2010). The difference between these two classes of mutations is also reflected in the performance of binding affinity change ($\Delta\Delta G$) predictors: both mCSM and HADDOCK (either raw or retrained) scores achieved much higher correlation on interface than on non-interface mutations. mCSM, which is based on a machine learning model with graph-based signatures based on more than 20 features including sequence profiles, performs much better than our simple HADDOCK score based on empirical energies. Only when trained separately against various experimental methods dataset does the HADDOCK score reach a reasonable performance.

As already shown in our previous work on contact-based prediction of binding affinity (Vangone and Bonvin, 2015), the experimental methods used for measuring binding affinity have a strong impact on the prediction performance: using the classification reported in DACUM, we demonstrated that both HADDOCK and mCSM show a great variability in performance depending on the experimental method, with correlation coefficients for interface mutations ranging from r=-0.446 (EMSA) to 0.918 (SFSP) for the raw HADDOCK score, and from r=-0.140 (ELISA) to 0.844 (FL) for mCSM. The performance of HADDOCK and mCSM on the ALL dataset (0.276 and 0.606, respectively) is close to the average performance calculated over all method datasets (0.28±0.34 and 0.50±0.25, respectively).

Even if only large datasets are considered (i.e. SPR, FL, SP, SFFL, IASP), the performance differences between datasets were still significantly large.

A better performance can be obtained by training on datasets for a specific method as demonstrated with the retrained HADDOCK predictor. But how realistic is that? In principle the impact of a mutation on the strength of an interaction should not depend on the experimental method used to measure it. The observed difference might rather reflect different accuracies of experimental measurement methods, next to of course all the various factors that can affect binding affinity measurements, like salt concentration, buffer, pH, temperature. For example, the failed training on CSPRIA, ELISA and EMSA datasets may suggest a low quality of the reported data. These large differences between datasets raises the questions of which datasets should ideally be used for training a new predictor and what is the reliability of various experimental methods. To illustrate this, we calculated the prediction performance of the trained HADDOCK score on the various methods datasets not used for training (Figure 3). The resulting correlation matrix clearly shows that predictors trained on some datasets seem to perform better overall. The prediction performance of each predictor over the other datasets (independent validation) is indicated on the right side of the matrix (boxed column). From these, SPR, FL, SFFL, SFSP, IAFL and RA seem to perform best (with an average r over the independent sets ranging from 0.28 to 0.37, possibly indicating that the corresponding data sets are more reliable. All together, these datasets represent 902 mutations, which is already a nice set for developing new predictors. Retraining a predictor on this subset using four-fold cross-validation results in an improved HADDOCK predictor performance of r=0.391±0.005 (cross-validated r) (see SI Table S3). The HADDOCK raw score performance in the same subset is r=0.363, indicating that our simple scoring function is already reasonably robust (optimization does not improve it much), but too simple for ΔΔG prediction compared to the more sophisticated mCSM method.

For the future research, it will thus be important to take the experimental methods and position of mutation into account when testing and training new binding affinity change predictors for protein complexes. The DACUM database, which is providing this information, is only a start. Ideally one would hope that a large, reliable and consistent set of ΔΔG measurements will become available in the future. For the time being we have to rely on a heterogeneous collection of data, obtained mainly in academic labs. The industry could play

here significant role by making data available since the experimental conditions for their measurements might be much more controlled and uniform than in an academic setting.

5. Acknowledgements

C. Geng and A. Vangone acknowledges financial support from the China Scholarship Council [grant number 201406220132], and Marie Skłodowska-Curie Individual Fellowship MSCA-IF-2015 [grant number BAP-659025], respectively.

6. Reference

Berliner, N., Teyra, J., Çolak, R., Lopez, S. G. and Kim, P. M. (2014) *PLoS ONE*,**9** e107353.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) *Nucl. Acids Res.*,**28** 235–242.

Brender, J. R. and Zhang, Y. (2015) *PLOS Comput Biol*,**11** e1004494.

Cyril Dominguez, Rolf Boelens, A. and Bonvin, A. M. J. J. (2003) *J. Am. Chem. Soc.*,**125** 1731–1737.

Davis, I. W. et al. (2007) *Nucl. Acids Res.*,**35** W375–W383.

Dehouck, Y., Kwasigroch, J. M., Rooman, M. and Gilis, D. (2013) *Nucl. Acids Res.*,**41** W333–W339.

Dourado, D. F. A. R. and Flores, S. C. (2014) *Proteins: Structure, Function, and Bioinformatics*,**82** 2681–2690.

Fernández-Recio, J., Totrov, M. and Abagyan, R. (2004) *Journal of Molecular Biology*,**335** 843–865.

Jorgensen, W. L. and Tirado-Rives, J. (1988) *J. Am. Chem. Soc.*,**110** 1657–1666.

Kastritis, P. L. and Bonvin, A. M. J. J. (2010) *J. Proteome Res.*,**9** 2216–2225.

Kastritis, P. L. and Bonvin, A. M. J. J. (2013) *Journal of the Royal Society Interface*,**10** 20120835–20120835.

Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M. J. J. and Janin, J. (2011) *Protein Science*,**20** 482–491.

Lensink M.F., Velankar S., Kryshtafovych A., Huang S. et al. In press (2016) *Proteins: Struc. Funct. & Bioinformatics.*

Levy, E. D. (2010) *Journal of Molecular Biology*,**403** 660–670.

Li, M., Petukh, M., Alexov, E. and Panchenko, A. R. (2014) *J. Chem. Theory Comput.*,**10** 1770–1780.

Moal, I. H. and Fernández-Recio, J. (2012) *Bioinformatics*,**28** 2600–2607.

Moal, I. H. and Fernández-Recio, J. (2013) *J. Chem. Theory Comput.*,**9** 3715–3727.

Pires, D. E. V., Ascher, D. B. and Blundell, T. L. (2014) *Bioinformatics*,**30** 335–342.

R Development Core Team. 2014. R: A language and environment for statistical computing (Foundation for statistical computing). Vienna, Austria.

Stites, W. E. (1997) *Chem. Rev.*,**97** 1233–1250.

van Zundert, G. C. P. et al. (2016) *Journal of Molecular Biology*,**428** 720–725.

Vangone, A. and Bonvin, A. M. (2015) *eLife Sciences*,**4** e07454.

Figure 1. Boxplots of experimental ΔΔG values in DACUM

(A) Interface and **(B)** non-interface mutations. The "ALL" dataset is composed of all interface or non-interface mutations, irrespective of the experimental method. All other datasets are named using the abbreviation of corresponding experimental methods (see Table 1).

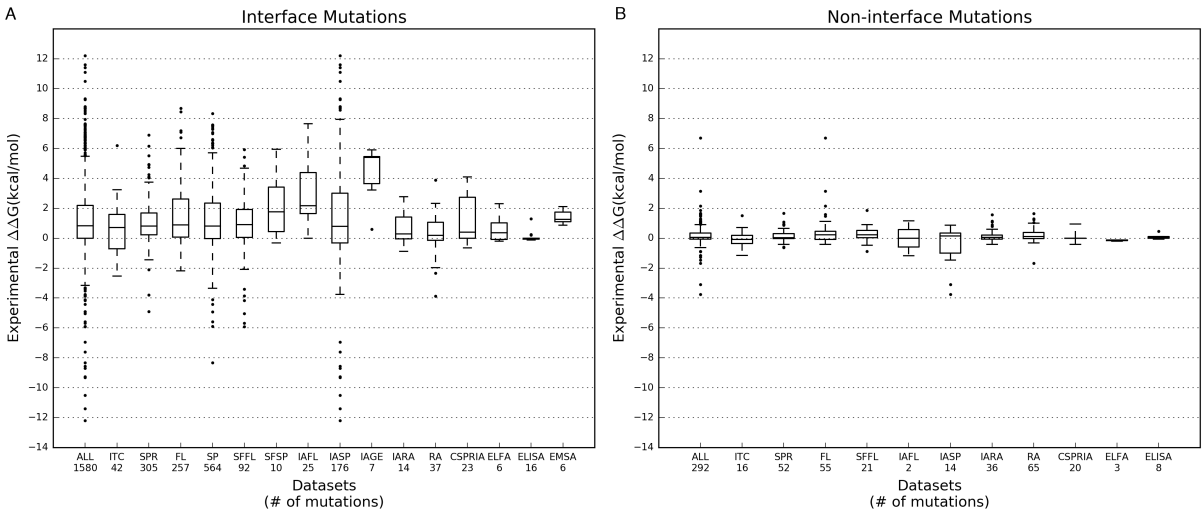
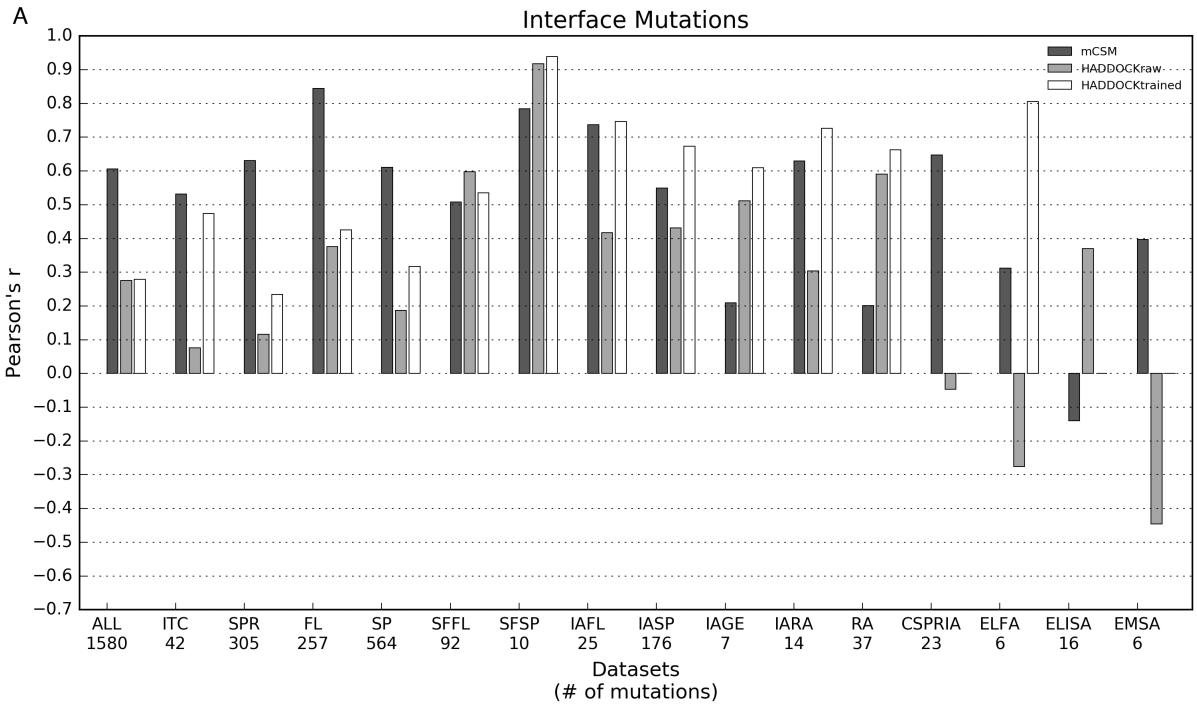


Figure 2. Performance of HADDOCK and mCSM on DACUM datasets

The Pearson's correlation coefficients of mCSM, raw HADDOCK score and trained HADDOCK ΔΔG predictors on interface mutations **(A)**, and that of mCSM on non-interface mutations **(B)**. Note that for CSPRIA, ELISA and EMSA datasets, there were no available trained HADDOCK predictors (see Table 2).



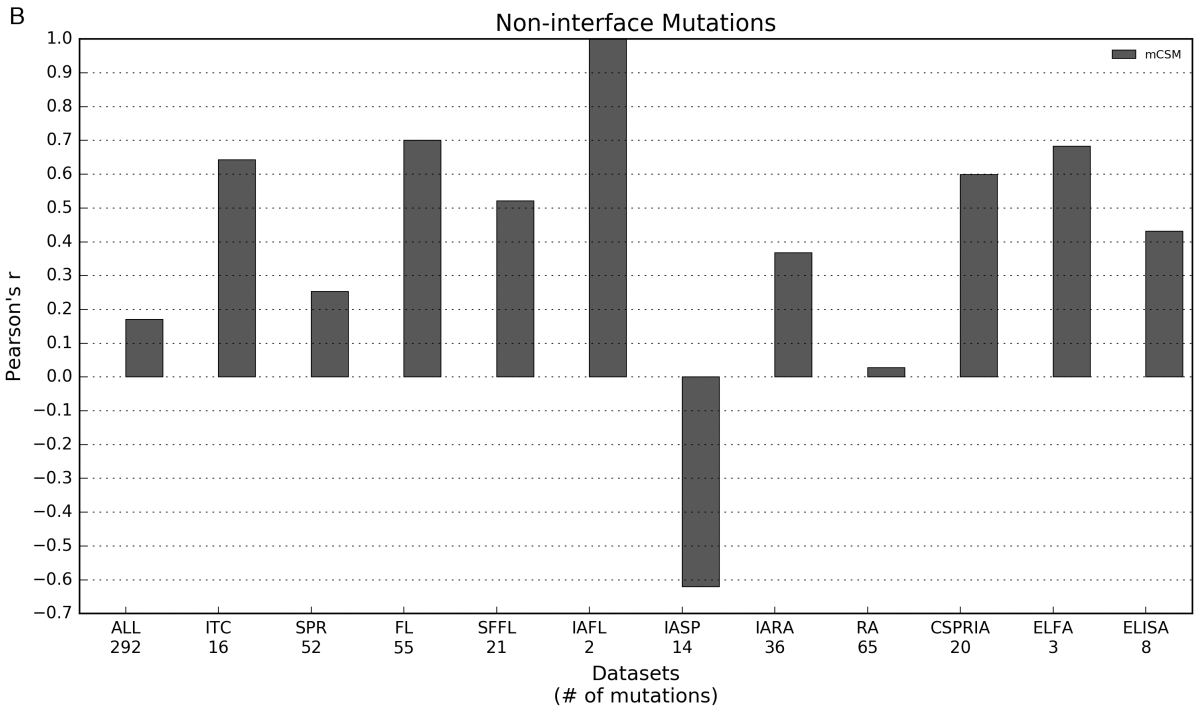


Figure 3. Performance of differently trained HADDOCK predictors on the various methods datasets

The performance of various trained HADDOCK ΔΔG predictors on interface mutations is shown for each independent methods dataset. The corresponding Pearson's correlation coefficients are color-coded following the scale shown on the right. Off-diagonal elements represent an independent validation, while the values on the diagonal give the correlation for the training dataset. The average performance (r), excluding the training dataset, is given on the right side of the matrix.

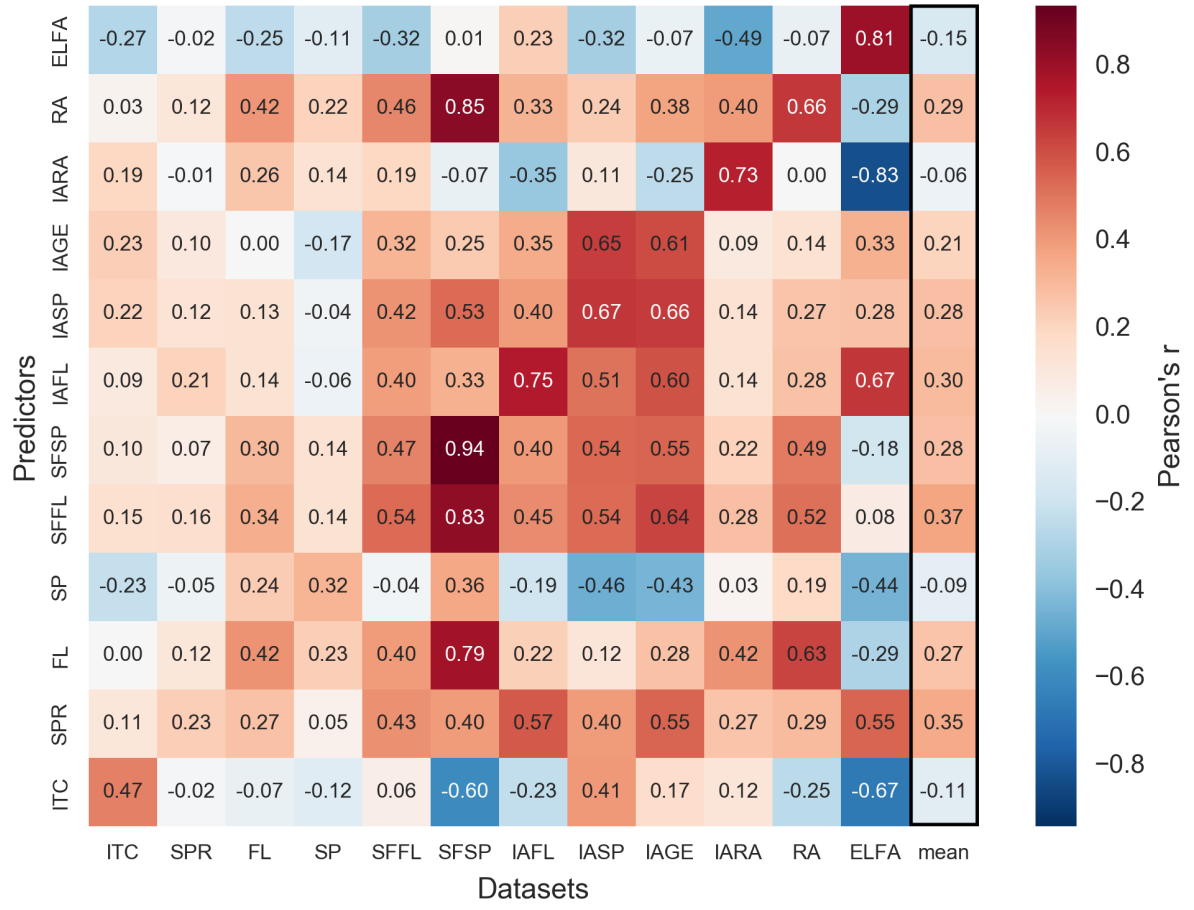


Table 1. Summary of experimental methods in DACUM

Abbreviation	Full name	Keywords of methods in reference[a]	Reference[b]
ITC	Isothermal Titration Calorimetry	ITC; Isothermal Titration Calorimetry;	11148036; 11420435; 15197281; 16177825; 16227441; 16867992; 17157249; 18471830; 18687868; 7629185; 8952503; 9092837;
SPR	Surface Plasmon Resonance	Surface Plasmon Resonance; SPR; BIAcore;	10772866; 10828942; 10864923; 10880432; 11123892; 11406576; 12006492; 12515535; 15064755; 15618400; 15791205; 16279951; 16300789; 16446445; 16923808; 17070843; 17976650; 18275829; 18319344; 18477456; 18687868; 19161338; 7504735; 7588629; 7654692; 8588944; 8703938; 8756685; 8962059; 8993317; 9050834; 9223641; 9367779; 9461077; 9500785; 9571026; 9579662; 9609690; 9878445; Water-mediated interaction at a protein-protein interface, Chemical Physics, Volume 307, Issues 2-3, 2004;
FL	Fluorescence	Fluorescence; Fluorimeter; Fluorimetry; Fluorometer; Spectrofluorimeter; Spectrofluorometer;	10452608; 11171964; 11278571; 12716886; 17475279; 18596201; 21642453; 7592655; 7716157; 8143850; 9047374; 9228059; 9632678; 9692956; 9788869; Stephen Ming-teh Lu, PhD Thesis, Purdue University, 2000;
SP	Spectroscopy	Spectroscopy; Spectrophotometer;	10338006; 11171964; 9047374;
SFFL	Stopped-Flow Fluorescence	Stopped-Flow Fluorescence; Stopped-Flow Fluorimeter; Stopped-Flow Fluorometer; Stopped-Flow Fluorescenceanisotropy; Excitation, Emission Fluorescence;	10065709; 10413501; 10876236; 10970748; 11136978; 2479414; 2742853; 7739054; 8494892; 9126847; 9425068; 9718299;
SFSP	Stopped-Flow Spectroscopy	Stopped-Flow Spectroscopy; Stopped-Flow Spectrophotometry;	9050852;
IAFL	Inhibition Assay Fluorescence	Inhibition Assay Fluorescence; Inhibition Assay Fluorescence Spectrophotometric;	11420435; 12515831; 20656696; 8507637; 9268350;
IASP	Inhibition Assay Spectroscopy	Inhibition Assay Spectroscopy; Inhibition Assay Spectrophotometer; Inhibition Assay Spectrophotometric; Inhibition Assay Spectrophotometricuv;	10065709; 10339415; 10691989; 1281426; 15284234; 15504027; 15865427; 16809340; 17405861; 18775544; 1992167; 7592720; 7683415; 7947796; 8063780; 8157652; 8276767; 8784199; 9048543; 9480775; 9761467; 9761468; 9891008;
IAGE	Inhibition Assay	Inhibition Assay	17138564;

	Gelelectrophoresis	Gelelectrophoresis;	
IARA	Inhibition Assay Radioactivity	Inhibition, ^{125}I-labeled, radioactivity; Competitive displacement assay, ^{125}I-labeled; Competitive binding assay, radioactively labeled;	10518943; 2034689; 8332602;
RA	Radioactivity	Radioactive subunit exchange; I^{125}-labeled;	18471830; 2402498; 2471267; 7529940;
CSPRIA	Competition Solid-Phase Radio-Immune Assay	Competition Solid-Phase Radio-Immune Assay;	1711212;
ELFA	Enzyme-Linked Functional Assay	Enzyme-Linked Functional Assay;	7756258;
ELISA	Enzyme-Linked Immunosorbent Assay	Enzyme-Linked Immunosorbent Assay;	7947809; 8312277; 9878445;
EMSA	Electrophoretic Mobility Shift Assay	Electrophoretic Mobility Shift Assay;	10984496;

a. The keywords of experimental methods occurred in references are reported here.

b. References using specific experimental methods are reported. The PubMed ID is given where available, otherwise the whole reference is provided.

Table 2. The parameters of predictors trained on different DACUM datasets using HADDOCK terms

	ALL	ITC	SPR	FL	SP	SFFL	SFSP	IAFL	IASP	IAGE	IARA	RA	ELFA
Intercept	1.083	0.263	0.922	1.267	1.037	0.570	0.884	1.797	0.691	4.576	0.736	0.112	1.052
$\mathbf{W_{vdw}}$	0.068	-0.104	0.063	0.133	0.074	0.125	0.110	0.317	0.101	-	-	0.176	-
$\mathbf{W_{elec}}$	0.015	-	0.006	0.008	-0.014	0.022	0.050	0.048	0.061	0.024	-0.016	0.017	-
$\mathbf{W_{desolv}}$	0.049	-0.063	-	0.063	0.076	0.060	0.219	-	0.060	-	-	0.107	-
$\mathbf{W_{BSA}}$	-	-0.014	0.002	-	-	-	-	0.020	-	-	-0.034	-	0.008
Pearson r	0.279	0.474	0.234	0.425	0.317	0.535	0.939	0.747	0.673	0.610	0.727	0.663	0.806

The trained HADDOCK ΔΔG predictors (Eq. 9) are named using the abbreviation of experimental methods. HADDOCK terms Evdw, Eelec, Edesolv and BSA were used for training, and corresponding weights are reported with the intercept. The HADDOCK terms evaluated as not relevant from the Akaike Information Criterion (AIC) evaluation are reported as "-". Note that AIC did not select any term as relevant for CSPRIA, ELISA and EMSA, which are therefore not shown in the table.