Received: 21 September 2018 | Revised: 10 December 2018 | Accepted: 11 December 2018

DOI: 10.1002/wcms.1410

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

ADVANCED REVIEW

Finding the ΔΔ*G* spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it?

Cunliang Geng | Li C. Xue | Jorge Roel-Touris | Alexandre M. J. J. Bonvin

Bijvoet Center for Biomolecular Research, Faculty of Science—Chemistry, Utrecht University, Utrecht, The Netherlands

Correspondence
Alexandre M. J. J. Bonvin, Bijvoet Center for Biomolecular Research, Faculty of Science—Chemistry, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.
Email: a.m.j.j.bonvin@uu.nl

Funding information
China Scholarship Council, Grant/Award Number: 201406220132; Horizon 2020 Framework Programme BioExcel, Grant/Award Number: 675728; Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 718.015.001; Netherlands eScience Center ASDI, Grant/Award Number: 027016G04

Predicting the structure and thermodynamics of protein–protein interactions (PPIs) are key to a proper understanding and modulation of their function. Since experimental methods might not be able to catch up with the fast growth of genomic data, computational alternatives are therefore required. We present here a review dealing with various aspects of predicting binding affinity changes upon mutations (ΔΔ*G*). We focus on predictors that consider three-dimensional structure information to estimate the impact of mutations on the binding affinity of a protein–protein complex, excluding the rigorous free energy perturbation methods. Training and evaluation, ΔΔ*G* databases, data selection, and existing ΔΔ*G* predictors are specially emphasized. We also establish the parallel with scoring functions used in docking since those share many similar PPI features with ΔΔG predictors. The field has seen a common evolution of ΔΔ*G* predictors and scoring functions over time, transforming from purely energetic functions to statistical energy-based and further to machine learning-based functions. As machine learning has come to age, limitations in terms of quantity, quality and variety of the available data become the bottlenecks for the future development of these computational methods. This can be alleviated by building infrastructures for data generation, collection and sharing. Further developments can be catalyzed by conducting community-wide blind challenges for method assessment.

This article is categorized under:
Structure and Mechanism > Molecular Structures
Structure and Mechanism > Computational Biochemistry and Biophysics
Molecular and Statistical Mechanics > Molecular Interactions

KEYWORDS
ΔΔ*G* prediction, binding affinity, machine learning, mutations, protein–protein interactions, scoring function

1 | INTRODUCTION

Protein–protein interactions (PPIs) are central to most biological functions and activities, such as signal transduction, immune response, etc. Three-dimensional (3D) information on the structure of protein complexes can provide the physical basis to a full understanding of PPIs. A variety of experimental methods have been developed to determine or characterize the structure of protein assemblies at different resolutions, ranging from X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM), capable of providing atomic resolution information, to cross-linked mass spectrometry

(XL-MS), hydrogen/deuterium exchange, mutagenesis and a variety of other biophysical and biochemical methods that provide partial information such as distances between specific amino acids or binding site mapping.[1–4] Experimental studies remain, however, time consuming and are limited by various technical challenges. As a consequence only ~6.5% of the known human interactome has structural information[5] (Data source: Interactome3D, version 2018_04). Furthermore, the rate of experimental structure determination lags far behind the discovery rate of PPIs. Hopefully, this might change in the coming years due to the revolution that has taken place in cryo-EM which is particularly well-suited to characterize large macromolecular assemblies.[6–8] Because of these limitations, there is a need for complementary computational modeling methods that allow populating the 3D space of PPIs. One such method that is able to provide structural detail of PPIs in a high-throughput way is computational docking[9–13] (Figure 1). One of the main challenges for docking is to identify and select native-like structural models from a typically large ensemble of models generated by docking, the so-called scoring problem. To solve this challenge, a lot of efforts have been devoted to the development of reliable scoring functions.[14,15]

Next to gaining structural insights, understanding the thermodynamics of PPIs is key to revealing their mechanism of action, understanding the effect of disease-related mutations and/or engineering new interactions. The most important thermodynamic information that tells us the strength of interactions between proteins is the binding affinity or binding free energy, ΔG. Changes in binding affinity caused by mutations (i.e., $\Delta\Delta G$) can show the impact of mutations on PPIs. Binding affinity can be determined through various experimental methods, which usually require laborious and expensive experimental work. The research field could benefit from the fast and reliable computational tools to predict binding affinity and its changes upon mutations (Figure 1). The problem of binding affinity prediction has been previously reviewed.[16–19] The impact of mutation on binding affinity can also be treated as a classification problem, known as hot-spot prediction in this case, which is not covered in this review (for review see References 20,21). Here, we focus on the high-throughput prediction of binding affinity changes upon mutations ($\Delta\Delta G$), excluding the rigorous physical methods to calculate $\Delta\Delta G$, which are usually computationally expensive, such as free energy perturbation and thermodynamic integration.[22,23]

In Section 2 we discuss aspects of training and validation, relevant databases, and give an overview of representative $\Delta\Delta G$ predictors and their important features. Since many of these features are also often used in scoring functions, we then shortly discuss and compare representative docking scoring functions in Section 3. We further discuss the importance of data quantity, quality, and variety in Section 4. And the importance of blind challenges in development of computational tools is discussed in Section 5. The final Section 6 discusses prospects for $\Delta\Delta G$ predictors and scoring functions.

2 | PREDICTING BINDING AFFINITY CHANGES IN PPIs

The binding affinity change caused by mutations (i.e., $\Delta\Delta G$) is defined as the difference of binding affinity between mutant and wildtype protein complexes. Various computational approaches have been developed over the years to predict $\Delta\Delta G$.

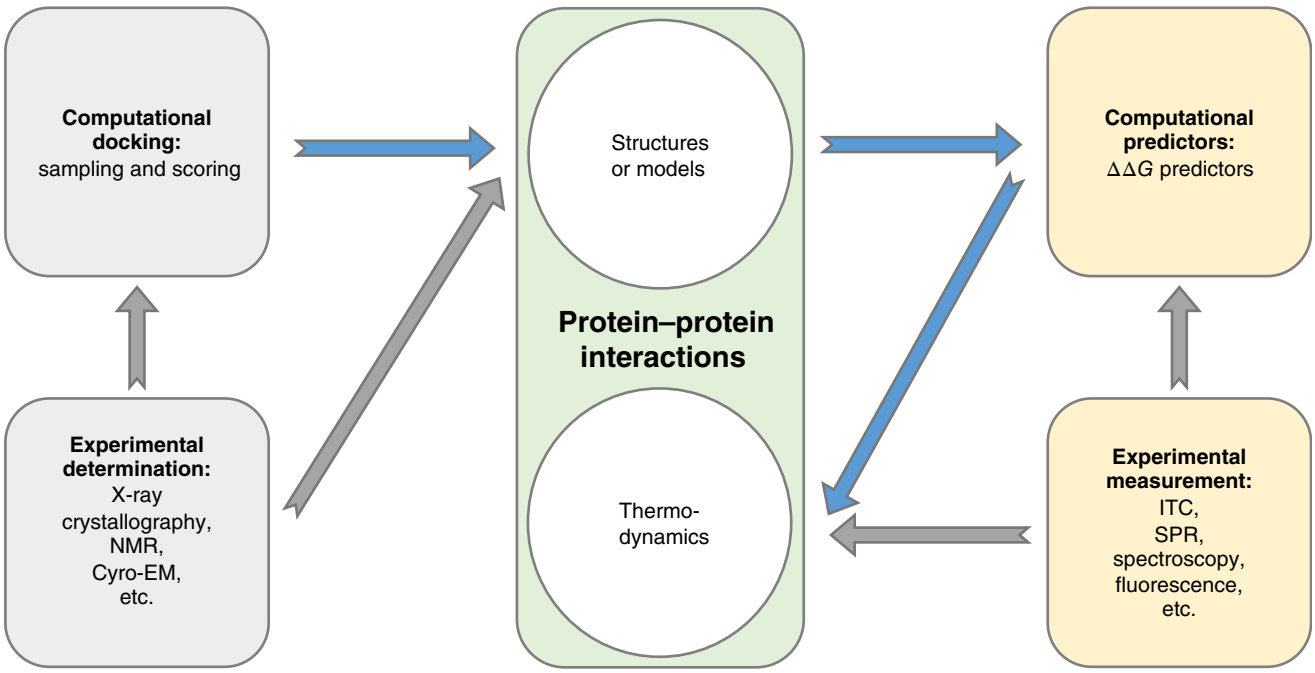

FIGURE 1 Approaches to study structural and thermodynamic features of protein–protein interactions

Considering the importance of 3D structure to provide the physical basis of PPIs, we focus here on the structure-based $\Delta\Delta G$ predictors, which base their predictions on experimental 3D structures or structural models of protein complexes.

2.1 | Training, validation, and test of $\Delta\Delta G$ predictors

The structure-based $\Delta\Delta G$ predictors in this review refer mainly to computational regression models used to quantitatively estimate the binding affinity changes upon mutations in protein–protein complexes. They are usually trained by optimizing a function based on various structural and sequence features of PPIs in order to reproduce experimental $\Delta\Delta G$ values. The form of the model can be a classical linear function, which represents $\Delta\Delta G$ as a sum of weighted features, or a machine learning-based nonlinear function, which has no predetermined form and infers the relationship between $\Delta\Delta G$ and features from experimental data.[24] Validation is necessary to get a properly optimized model. For this, usually N-fold cross validation (CV) is used.[25] In N-fold CV, a set of data is randomly and evenly split into N groups; $N - 1$ groups are used for training and the left-out one for validation. This process is repeated N times to evaluate the model on each one of the N groups. Additionally, leave-one-group-out CV (LOGCV) is useful to test a model's generalizability,[25] by classifying the data into different groups based on protein properties (e.g., complex type, protein family, sequence identity, structure similarity, etc.). The training/validation process is similar to N-fold CV. After training and validation, a final model is generated, usually based on all data. To blindly test the performance of the final model, an additional set of data not used in the training and validation process is needed.

To evaluate the performance of a $\Delta\Delta G$ predictor, root mean square error (RMSE), and Pearson's correlation coefficient (PCC) are commonly used:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_1^n \left(\Delta\Delta G_{\mathrm{exp}} - \Delta\Delta G_{\mathrm{pred}}\right)^2}{n}}$$

$$\mathrm{PCC} = \frac{cov\left(\Delta\Delta G_{\mathrm{exp}}, \Delta\Delta G_{\mathrm{pred}}\right)}{\sigma_{\Delta\Delta G_{\mathrm{exp}}} \cdot \sigma_{\Delta\Delta G_{\mathrm{pred}}}}$$

where n represents the number of mutations, cov is the covariance, and σ is the variance.

2.2 | $\Delta\Delta G$ databases

Curating and integrating structural and thermodynamic data of protein–protein complexes into databases is fundamental to promote the development of computational tools. The first database that combined structures and $\Delta\Delta G$ values is, to our knowledge, the Alanine Scanning Energetics database[26] (ASEdb) published in 2001 (Table 1). ASEdb contains experimentally determined binding affinities for single alanine mutations in protein–protein, protein–nucleic acid, and protein–small molecule interactions. Only a small part of these has 3D structural information for the associated complexes. In 2006, the Protein–protein Interactions Thermodynamic database[27] (PINT) was released. It contains multiple thermodynamic parameters including $\Delta\Delta G$, dissociation constants K_D, enthalpy change ΔH, heat capacity change ΔCp and so on, along with structural information. In 2012, the Structural Kinetic and Energetic database of Mutant Protein Interactions[28] (SKEMPI) came out. It has been the most frequently used database for developing $\Delta\Delta G$ predictors since then. SKEMPI version 1.1 collects >3,000 $\Delta\Delta G$ measurements for single and multiple mutations of 85 diverse protein–protein complexes with available 3D structures. It contains additional thermodynamic parameters when available in the literature. Since SKEMPI 1.1 did not provide information about the experimental methods used to determine the $\Delta\Delta G$ values, DACUM[29] (Database of binding Affinity Change Upon Mutations in protein complexes) was created as a subset of SKEMPI focusing on only single mutations. Taking into consideration the experimental techniques used for measuring $\Delta\Delta G$ revealed that those have a strong impact on the achievable accuracy of $\Delta\Delta G$ predictors,[29] underscoring the importance of properly considering and carefully choosing experimental data in terms of experimental methods when developing a $\Delta\Delta G$ predictor. Next to general $\Delta\Delta G$ databases for protein–protein complexes, the Antibody-Bind database[33] (AB-Bind) contains >1,000 mutations with experimental $\Delta\Delta G$ values for 32 antibody–antigen complexes and the Altered TCR Ligand Affinities and Structures database[34] (ATLAS) comprises binding affinities for wildtype and mutant TCR-pMHC complexes. In 2017, by integrating the previous databases with new information from literature, two new $\Delta\Delta G$ databases were released: PROXiMATE[30] (PROtein–protein compleX MutAtion ThErmodynamics) and dbMPIKT[31] (a kinetic and thermodynamic database of mutant protein interactions). Finally, SKEMPI 2.0[32] was released in 2018, which represents a major update of the first version, effectively doubling its size. It was obtained by combining several databases including SKEMPI 1.1, AB-Bind, PROXiMATE and dbMPIKT, and adding new manually curated data from literature, resulting in binding thermodynamics data for a total of 7,085 mutations, including kinetic data (K_{on}, K_{off}) when available. SKEMPI 2.0 contains ~3,000 single point alanine mutations, ~2,000 single point nonalanine mutations and

TABLE 1 List of representative $\Delta\Delta G$ databases

Type	Name	Year published	Data accessibility	Description
Comprehensive	ASEdb[26]	2001	http://nic.ucsf.edu/asedb/	620 binding affinity data cross-referenced to dimer structures in the PDB, covering 26 systems Only single alanine mutations Not maintained any more
	PINT[27]	2006	http://www.bioinfodatabase.com/pint/	699 binding affinity data cross-referenced to dimer structures in PDB, covering 32 systems
	SKEMPI[28]	2012	https://life.bsc.es/pid/mutation_database/index.html	3,047 $\Delta\Delta G$ measurements of 2,792 single or multiple point mutations for 158 structures of 85 systems Integrates data from ASEdb and PINT
	DACUM[29]	2016	https://github.com/haddocking/dacum	Subset of SKEMPI 1.1 for single point mutations (1872 entries for 81 systems) with additional information on experimental methods used to measure the binding affinity
	PROXiMATE[30]	2017	https://www.iitm.ac.in/bioinfo/proximate/	6,327 mutations for 176 complexes (checked on August 15, 2018) Includes data from ASEdb, PINT, SKEMPI 1.1 and AB-Bind
	dbMPIKT[31]	2017	See Reference 31	5,291 mutations Integrates data from SKEMPI 1.1 and AB-Bind
	SKEMPI 2.0[32]	2018	https://life.bsc.es/pid/skempi2	7,085 mutations for 345 structures of 237 systems Integrates data from SKEMPI 1.1, AB-bind, PROXiMATE and dbMPIKT To date the largest $\Delta\Delta G$ database
Complex-specific	AB-Bind[33]	2015	See Reference 33	1,101 mutations for 32 antibody complexes
	ATLAS[34]	2017	http://atlas.wenglab.org/web/index.php	694 affinity data for wildtype and mutant TCR-pMHC complexes of human and mouse
Low quality	Modeled structures	2016	See Reference 35	846 mutations for 15 complexes, with structures generated through modeling and fitting
	Deep sequencing	2016	See Reference 36	5,748 single point mutations of type I dockerin-cohesin complexes

~2,000 multiple mutations on various types of protein complexes, such as protease-inhibitor, antibody–antigen, TRC-pMHC complexes, etc. In addition, unlike SKEMPI 1.1, the experimental methods used for measuring the binding affinities are now reported. About 60% of the data have been measured by Surface Plasmon Resonance (SPR), Fluorescence, and Isothermal Titration Calorimetry (ITC) methods.

Besides the databases that contains high-resolution structures and binding data measured by traditional experimental methods, several interesting databases provide structural models and deep sequencing-based binding data. Dourado and Flores[35] have built a $\Delta\Delta G$ dataset based on both homology models and fitted models generated by fitting into low-resolution Cryo-EM density map. Deep sequencing is a high-throughput experimental method that can detect mutations, and provides estimates of experimental binding affinity based on the enrichment data.[37–39] The Critical Assessment of Prediction of Interactions (CAPRI), a community-wide experiment on the comparative evaluation of protein–protein docking for structure prediction, has run a $\Delta\Delta G$ prediction challenge based on experimental data from deep sequencing.[40] Deep sequencing was also exploited for two specific complexes, type I dockerin-cohesin[36] and TCR-pepMHC,[41] to study the impact of mutations and train $\Delta\Delta G$ predictors.

2.3 | Data selection for training $\Delta\Delta G$ predictors

To develop a $\Delta\Delta G$ predictor, the first step is usually to select a proper dataset for training. The data selection usually considers four aspects, namely the type of the complex, the available 3D structural information, the experimental $\Delta\Delta G$ values, and the type of mutations:

- *Complex type*: A complex may contain two chains (dimer) or multiple chains (multimer). Some predictors can only handle dimers.[42,43] The complexes are also often clustered based on their sequence identity to remove redundancy. Usually, the complex with the largest number of mutations from a cluster is kept.[44] In addition, some methods concentrate on specific types of complexes such as antibody–antigen complexes[33] or TCR-pMHC complexes[45] to develop complex-specific $\Delta\Delta G$ predictors.
- *Structural information*: The quality of a 3D structure is a key point in selecting a dataset, especially since many features are calculated from the 3D structure. For this reason, the highest resolution structures are often selected. Databases, such

as SKEMPI,[28,32] provide the Protein Data Bank (PDB)[46] entry of the highest resolution wildtype complex if more than one structure has been deposited. In principle, also lower-resolution structures or models could be explored to train and test $\Delta\Delta G$ predictors. Some attempts along those directions have been reported (see below).[35,43,44]

- *The experimental $\Delta\Delta G$ values*: Their quality is very important. The different accuracies and limitations of the experimental methods used to obtain $\Delta\Delta G$ values have been shown to impact the performance of $\Delta\Delta G$ predictors.[29] Multiple experimental $\Delta\Delta G$ values from different methods or conditions might be available for the same mutation. Some work have used all these $\Delta\Delta G$ values for training,[47] other took their average,[42,48–51] or selected the values originating from what the authors consider the most reliable experimental methods.[29,49,51] Mutations with binding affinities out of the experimental detection range (e.g., with only an upper limit, or absence of binding are reported without exact $\Delta\Delta G$ value) should be excluded for training and testing regression-based $\Delta\Delta G$ predictors. The new SKEMPI 2.0 database reports 440 such mutations.
- *Type of mutation*: As for the mutation itself, its location, type, frequency and other details (e.g., is it a single mutation or does it belong to a multiple mutation measurement?) should be considered. For example, mutations in the core of a protein, or outside the interface of a complex[29,43,44,49] and mutations from or to Proline[49] which are often causing conformational changes difficult to predict could be discarded. Some methods only concentrate on alanine mutations.[52,53] Only single point mutations,[44,47,48,51,54] multiple point mutations,[54] or both types[42,49,53] could be selected. Also, bias in the type of mutations and their frequencies could be corrected when selecting mutations in order to obtain a balanced dataset.[49,55]

2.4 | Overview of representative $\Delta\Delta G$ predictors

Many structure-based $\Delta\Delta G$ predictors have been published in the past two decades (Figure 2, Table 2). In general, these predictors can be classified into classical linear functions and machine learning-based methods. Classical linear functions typically rely on physical energies and/or statistical potentials with weights optimized to reproduce the experimental data. In contrast, machine learning-based functions might use a variety of information from structure, energy, evolution, etc. to estimate $\Delta\Delta G$.

The earliest $\Delta\Delta G$ predictors mostly used physical energies as features to describe PPIs. FoldX[56,57] is a representative of such predictors. It is still wildly used for predicting mutation effects. It is based on physical energies such as van der Waals and electrostatic energies, with additional hydrogen bond and solvation contributions. To model mutations, FoldX uses a rotamer approach, only allowing conformational changes of sidechains and keeping the backbone fixed. In the same year that FoldX was published, a $\Delta\Delta G$ predictor based on Rosetta was published,[58] relying on a linear combination of physical energies

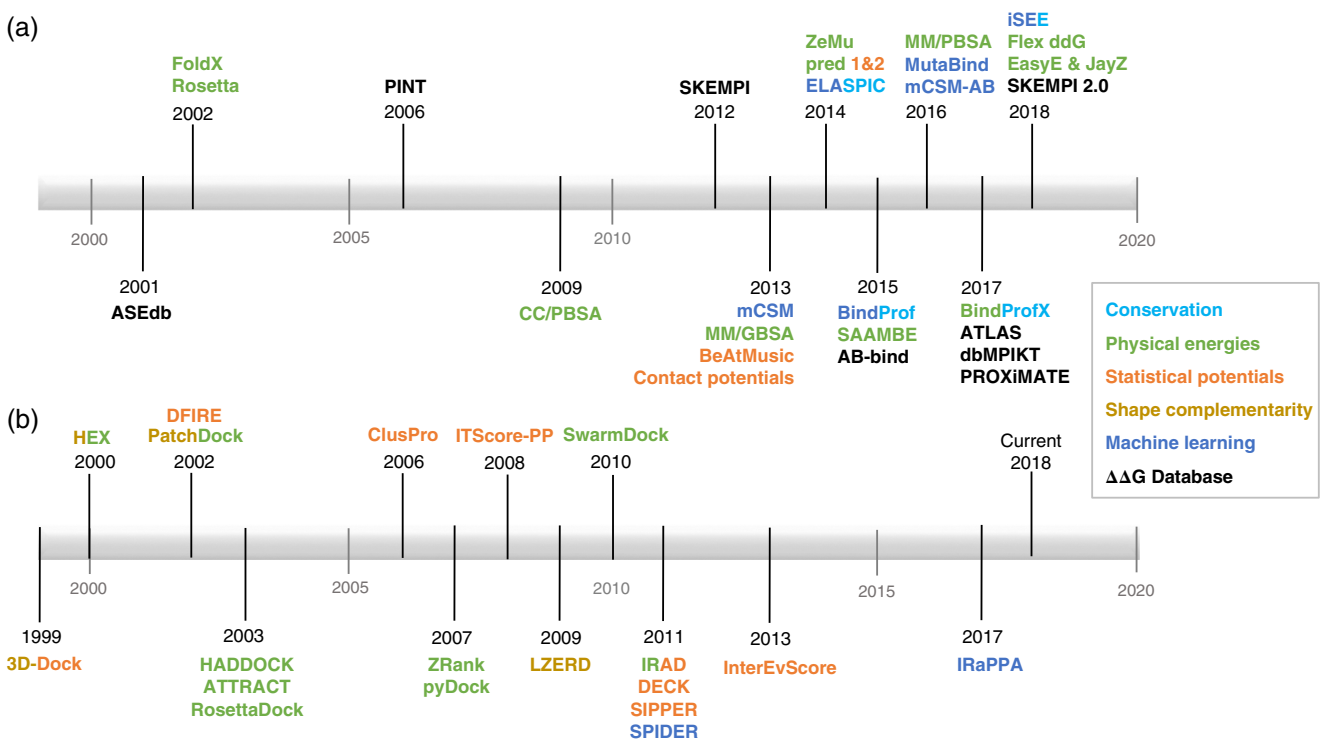

FIGURE 2 Timeline of representative $\Delta\Delta G$ predictors and databases (a) and scoring functions (b) with the color coding representing the type of features used (in color), or highlighting the databases (in black). This timeline is showing selected, representative methods and databases, and is by no means as an extensive representation of all published work

TABLE 2 List of representative $\Delta\Delta G$ predictors

Type	Name	Webserver/software	Description
Physical/empirical energies based	FoldX[56,57]	http://foldxsuite.crg.eu/	Models mutations using rotamer approach
	Rosetta[58]	NA	Models mutations using rotamer approach
	CC/PBSA[59]	NA	Uses structural ensemble to model flexibility of backbone and sidechain
	ZEMu[35,55]	https://simtk.org/projects/rnatoolbox	Uses multiscale approach to model flexibility of mutation region. $\Delta\Delta G$ predictions are based on FoldX
	Flex ddG[53]	NA	Samples structure flexibility with Rosetta "backrub". $\Delta\Delta G$ predictions are based on Rosetta energy function
	EasyE and JayZ[60]	NA	Guarantees identification of conformation of global minimum energy
	Beard et al.[61]	NA	Directly applies MM/GBSA to predict ΔΔG of single point mutations
	Li et al. Pred 2[54]	NA	Combines MM/PBSA with predictions from FoldX and BeAtMuSiC
	SAAMBE[50,62]	http://compbio.clemson.edu/saambe_webserver/	Combines MM/PBSA and statistical potential. Using amino acid specific dielectric constants to mimic mutations flexibility
	Simões et al.[52]	NA	Uses MM/PBSA and amino acid-specific dielectric constants to mimic mutations flexibility
	BindProfX[43]	https://zhanglab.ccmb.med.umich.edu/bindprofx/	Combines interface profile (conservation) with FoldX prediction
	Borrman et al.[34]	NA	TCR-pMHC complex-specific $\Delta\Delta G$ predictor
Statistical potentials based	BeAtMuSiC[48]	http://babylone.ulb.ac.be/beatmusic/	Uses coarse-grained model
	Contact potentials[49]	NA	Atomic and residue-level contacts
Machine learning based	mCSM[47]	http://biosig.unimelb.edu.au/mcsm/	Uses atom distance patterns with Gaussian process regression
	mCSM-AB[63]	http://biosig.unimelb.edu.au/mcsm_ab/	Antibody-specific $\Delta\Delta G$ predictor
	ELASPIC[44,64]	http://elaspic.kimlab.org/	Uses 75 sequence, energy and molecular features with decision trees
	BindProf[42]	https://zhanglab.ccmb.med.umich.edu/bindprof/	Combines interface profiles with FoldX energies through random forest
	MutaBind[51]	https://www.ncbi.nlm.nih.gov/research/mutabind/	Uses various structure, energy and conservation features with both multiple linear regression and random forest
	iSEE[65]	https://github.com/haddocking/isee	Combines 31 structure, energy and evolutionary conservation features with random forest

dominated by Lennard Jones interactions, solvation interactions, and hydrogen bonding. It also uses a rotamer approach and was limited to alanine mutations. CC/PBSA (Concoord/Poisson–Boltzmann surface area)[59] published in 2009, was one of the first $\Delta\Delta G$ predictors to use structural ensembles to explicitly describe conformational changes of both backbone and sidechains. It estimates $\Delta\Delta G$ by using averaged physical energies on all minimized structures of an ensemble. The method is, however, computationally expensive.[55]

Besides physical energies, statistical potentials can also be used to estimate $\Delta\Delta G$. These are typically extracted from experimental structures and therefore highly dependent on the availability of large and various experimental training data. The SKEMPI database[28] does provide such a large dataset. The authors of SKEMPI developed a statistical energy-based $\Delta\Delta G$ predictor based on intermolecular contact potentials[49] derived from atomic and residue-level contacts with and without consideration of data bias to specific families, complexes and residue types. Residue-level contact potentials were shown to perform slightly better than atomic ones achieving a PCC of 0.73 on 1949 single and multiple point mutations of SKEMPI. When data bias was considered lower PCC values were obtained.[49] BeAtMuSiC[48] is another statistical energy-based $\Delta\Delta G$ predictor based on a coarse-grained model. It achieves a PCC of 0.68 and RMSE of 1.19 kcal/mol on the single point mutations of SKEMPI 1.1.

Accounting for mutation-induced conformational changes in protein complexes is very important for the energy-based $\Delta\Delta G$ predictors, because the calculated energies are rather sensitive to details of the structures used. Early $\Delta\Delta G$ predictors such as FoldX[56,57] and Rosetta[58] model sidechain conformational changes using rotamer libraries, but these approaches cannot guarantee the identification of the lowest energy conformation. EasyE[60] could identify the conformation of global minimum energy with high confidence with the guaranteed Cost Function Network algorithm, resulting in an improved performance over FoldX. Such rotamer approaches, however, do not consider backbone flexibility. Since the early predictors, many

attempts have been made to explore structural flexibility in order to improve the performance of $\Delta\Delta G$ predictors. ZEMu[55] (Zone Equilibration of Mutants) combines structural flexibility with FoldX $\Delta\Delta G$ calculations using a multiscale approach. It limits the conformational sampling to a small region around the mutation site and leaves distant regions unperturbed, based on the assumption that mutations do not induce changes in global tertiary structure.[40,66] This led to an improvement in PCC from 0.49 for FoldX to 0.62 for ZEMu on 1,254 single and multiple point mutations of SKEMPI. Flex ddG[53] is another $\Delta\Delta G$ predictor that models conformational changes using Rosetta's Talaris energy function.[67–69] Unlike ZEMu that only samples a small region around the mutation site, Flex ddG considers the entire structure as flexible by generating an ensemble of conformations with the Rosetta "backrub"[70] protocol using torsion angle minimization and sidechain repacking. It estimates $\Delta\Delta G$ from the averaged Rosetta all-atom energies across the ensemble. Flex ddG showed considerable improvements of predictive performance over ZEMu on mutations such as small-to-large mutations (PCC from 0.48 to 0.65), multiple nonalanine mutations (PCC from 0.53 to 0.63), and mutations in antibody–antigen interfaces (PCC from 0.54 to 0.61).

Another group of classical linear $\Delta\Delta G$ predictors are based on Molecular Mechanics/Poisson−Boltzmann Surface Area (MM/PBSA) or Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) methods.[50,52,54,61] These methods typically use ensembles of conformations from molecular dynamics simulations, combining the molecular mechanics energies with solvation energies based on a continuum representation of the solvent (PBSA or GBSA schemes). These, alone or combined with other features, can be used to estimate $\Delta\Delta G$. Beard et al.[61] directly applied MM/GBSA energies to calculate $\Delta\Delta G$ of single point mutations observing statistically significant correlations for various complexes. Li et al.[54] applied a MM/PBSA method to predict $\Delta\Delta G$, and also combined it with FoldX and BeAtMuSiC into a multiple linear regression (MLR) model. Both standalone and combined $\Delta\Delta G$ predictors showed significant improvements in predictive performance (PCC: 0.69) over FoldX (PCC: 0.47) and BeAtMuSiC (PCC: 0.52). Likewise, SAAMBE[50] combines MM/PBSA with statistical terms derived from physicochemical properties of protein complexes, including residue entropy, hydrophobicity, hydrogen bonds, interface area and normalized changes of interface area. Moreover, it uses amino acid-specific dielectric constants to implicitly account for conformational changes induced by mutations. A similar strategy was also applied in another alanine mutations only $\Delta\Delta G$ predictor.[52] SAAMBE showed higher predictive accuracy than FoldX and BeAtMuSiC on different sets of single point mutations, such as large-to-small mutations, alanine mutations and mutations on different locations. These MM/PBSA- or MM/GBSA-based methods demonstrate a rather good prediction performance at a reasonable computational cost (not including the time required to generate the ensembles by MD simulations).

Machine learning methods have also been explored to predict $\Delta\Delta G$. Due to their data-driven nature, these highly rely on large and various experimental datasets. Their development has been catalyzed by the availability of large mutation databases (see above). One of the earliest such $\Delta\Delta G$ predictors is mCSM.[47] It mainly uses atomic distance patterns to represent the environment of the mutation site and was trained with a Gaussian process regression algorithm to estimate $\Delta\Delta G$, achieving a high performance (PCC: 0.801, RMSE: 1.251 kcal/mol) on all 2,317 single point mutations of SKEMPI 1.1 with 10-fold CVs. It also shows good performance (PCC: 0.56, RMSE: 1.38 kcal/mol) when blindly tested on a low redundancy subset (350 mutations) of SKEMPI 1.1. Notably, mCSM does not require modeling the mutation since it only uses the wildtype structure, representing the mutation by the induced changes of atom type counts.

Ensemble learning approaches such as decision trees and Random Forest (RF) are becoming popular in machine learning-based $\Delta\Delta G$ predictors. An example using the decision trees algorithm is ELASPIC.[44,64] It was developed using the Stochastic Gradient Boosting of Decision Trees (SGB-DT) algorithm to integrate a large and diverse set of features (75 in total). These include sequence (such as SIFT score[71]), energy (such as FoldX energy terms and $\Delta\Delta G$), and molecular features (such as solvent accessible surface area, hydrophobic and hydrophilic contributions). ELASPIC achieved a PCC of 0.75 and RMSE of 1.2638 kcal/mol on a subset of SKEMPI selected by applying a 90% sequence identity redundancy reduction, surpassing FoldX (PCC: 0.44) and BeAtMuSiC (PCC: 0.53) by a considerable margin. The RF algorithm has been most frequently used in $\Delta\Delta G$ predictors (such as BindProf,[42] iSEE[65], and MutaBind[51]). BindProf[42] constructs an interface binding profile score from an aligned ensemble of structurally similar interfaces, representing residue conservation. As a standalone feature, the interface profile score had a predictive accuracy similar to energy-based predictors such as FoldX. By combining interface profiles with atomic and residue level energies (e.g., FoldX energies) using a RF algorithm, BindProf got a very high CV performance (PCC 0.83) on the dimeric complexes of SKEMPI. iSEE[65] is another predictor based on a limited number of interface structural, evolution and energy-based features (31 in total). It describes the evolutionary conservation of mutation position by a Position Specific Scoring Matrix generated from multiple sequence alignment. Structural and energetic features are obtained from a short refinement of the complex with the HADDOCK[72,73] docking software. iSEE achieved a high CV performance (PCC: 0.80, RMSE: 1.41 kcal/mol) on single point mutations of dimeric complexes from SKEMPI 1.1. It shows competitive predictive performance compared to other classical and machine learning based $\Delta\Delta G$ predictors, such as FoldX,[56] CC/PBSA,[59] BeAtMuSiC,[48] ZeMu,[55] MM/PBSA,[54] mCSM,[47] and BindProfX.[43] Another RF-based $\Delta\Delta G$ predictor is MutaBind.[51] It combines RF with a MLR algorithm to integrate a variety of energetic, structural and conservation features

with a fast sidechain optimization. Its final $\Delta\Delta G$ prediction is the average of the predicted $\Delta\Delta G$ values generated by the RF and MLR algorithms. MutaBind achieved a PCC of 0.68 with Leave One Complex Out Cross Validation (LOCOCV) on the single point mutations of SKEMPI 1.1, outperforming FoldX and BeAtMuSiC (PCC: 0.40 and 0.39, respectively). In addition, MutaBind was also specifically trained on the protease-inhibitor complexes of SKEMPI, achieving a PCC of 0.76 with LOCOCV and again surpassing FoldX (0.40) and BeAtMuSiC (0.44).

Another interesting $\Delta\Delta G$ predictor is BindProfX.[43] Like BindProf,[42] BindProfX also uses the structure interface profile, which represents the conservation of interface residues. BindProfX, however, estimates $\Delta\Delta G$ using amino acid probabilities from the Boltzmann distribution, while BindProf uses log-odd likelihood derived scores. In particular, pseudo-counts were introduced in BindProfX to offset the limit of data samples for interface multiple structure alignment. The Boltzmann probability with pseudo-counts used in BindProfX shows better predictive correlation than not only the log-odd likelihood score but also FoldX and BeAtMuSiC on single or multiple point mutations. By combining Boltzmann probability and FoldX $\Delta\Delta G$ with a simple linear function, BindProfX shows a further improved PCC of 0.738 on 1,131 single mutations of SKEMPI.

2.5 | Complex-specific $\Delta\Delta G$ predictors

Most $\Delta\Delta G$ predictors have been developed to predict $\Delta\Delta G$ for various types of protein–protein complexes, which makes them generally applicable. Only few have been tailored to specific types of complexes. A main reason for this might be the lack of large amounts of experimental data for specific complexes. To increase the amount of relevant data for immune protein complexes, Sirin et al.[33] curated the AB-Bind database for antibody complexes and Borrman et al.[34] built the ATLAS database for TCR-pMHC complexes. The general $\Delta\Delta G$ predictors, such as FoldX and mCSM, were tested on AB-Bind data, showing low predictive performance (PCC: 0.34 and 0.35 for FoldX and mCSM, respectively).[33] By using the same types of features as mCSM,[47] Pires and Ascher[63] trained an antibody antigen-specific $\Delta\Delta G$ predictor, called mCSM-AB, which showed an improved PCC of 0.53 with 10-fold CV on the AB-Bind data. Borrman et al.[34] trained a specific $\Delta\Delta G$ predictor using energies and linear regression which achieved a PCC of 0.63 on the ALTAS database. While these complex type-specific predictors did show improved prediction performance, they did not achieve the prediction accuracy of predictors trained on general datasets. Developments are hampered by the availability of large, complex-specific datasets.

2.6 | Are lower-quality structural models good enough?

The availability of experimental structures for PPIs is rather limited.[5] This, however, can be largely expanded by computational modeling techniques.[5] In order to predict the impact of mutations at an interactome level, it is necessary to develop and/or test $\Delta\Delta G$ predictors that can also handle modeled structures. Some attempts have been reported, for example, for ELASPIC,[44,64] ZEMu[35,55], and BindProfX.[43]

In ELASPIC,[44,64] homology modeling was used to generate models for 2061 mutations of SKEMPI. A new $\Delta\Delta G$ predictor was then trained on these modeled structures, since it was found that features displayed different distributions between modeled and experimental structures. The new ELASPIC $\Delta\Delta G$ predictor achieved a PCC of 0.4 on the modeled structures, which is much lower than what was achieved (0.75) on experimental ones, but still better than FoldX alone (PCC: 0.27).

For ZEMu,[35] a validation dataset with 846 single and multiple point mutations in 11 models of protein complexes was created. Notably, not only homology modeling (with sequence identities ranging between 44% and 51%) but also fitting into Cryo-EM density maps was used to generate structural models of the complexes. ZEMu[35,55] achieved a PCC of 0.34 and RMSE of 1.54 for the set of 846 mutations. On a subset of 558 mutations for which experimental structures were available, ZEMu achieved a lower performance on modeled (PCC: 0.38, RMSE: 1.76) versus experimental complexes (PCC: 0.61, RMSE: 1.58), although it did outperform FoldX (PCC: 0.17, RMSE: 2.00). A similar predictive performance was observed on both homology models and cryo-EM-fitted structures.[35]

BindProfX[43] was tested on a subset (104 mutations) of the ZEMu low-resolution structure models for dimeric complexes, achieving a PCC of 0.454 and outperforming ZEMu (PCC 0.118).[43]

These attempts to use structural models for $\Delta\Delta G$ prediction all demonstrate that the quality of the models is crucial to achieve a good predictive performance since none of the tested/developed predictors achieved accuracy (PCC) above 0.5. This is significantly lower than the reported performances on experimental structures. There is thus still plenty of room for further improvement. This will require both better computational modeling and optimization techniques and more robust $\Delta\Delta G$ predictors.

2.7 | Important features

Knowledge of the most important features used to predict the impact of a mutation on the binding affinity of a complex should contribute to our understanding of PPI mechanisms and stimulate the further development of related computational tools. For

classical linear $\Delta\Delta G$ predictors, the weights of features in the linear function usually reflect their relative importance. In contrast, it is more difficult for machine learning methods to extract the importance of a specific feature due to their nonlinear form. Ensemble learning approaches such as decision trees and RF that can rank features by their contributions to the final prediction. For example, in decision trees and RF, the information gain or error decrease when selecting a given feature as the spitting node can be used to estimate the importance of this feature. The more information is gained, or the error is decreased, the more important that particular feature is. The importance of all features is then ranked by their total information gain or error decreased. Several $\Delta\Delta G$ predictors based on such approaches have reported the relative importance of the features used. For example, in ELASPIC,[44,64] which uses the SGB-DT algorithm, the FoldX $\Delta\Delta G$ was the most important feature, followed by other energy terms such as sidechain entropy, energy clashes and solvation of mutated residues. The sequence conservation (SIFT score[71]) was also highly important. Similarly, in the analysis of BindProf features with RF, FoldX $\Delta\Delta G$ and interface profile were the two most important features. The RF-based iSEE[65] predictor highlights the dominant role of residues conservation in $\Delta\Delta G$ prediction, followed by the changes of intermolecular electrostatic energy and changes of buried surface area. Taken together, these analyses underline the importance of intermolecular energies and residue conservation for estimating mutation effects on binding affinity of protein complexes.

3 | A PARALLEL WITH SCORING FUNCTIONS

$\Delta\Delta G$ predictors and scoring function used in docking do share many features although the questions they are trying to answer are different. Scoring in the context of docking is defined as the evaluation and selection of near-native poses out of the typically large pool of models generated by protein–protein docking algorithms. Scoring and dealing with conformational changes are still the two main challenges in the docking field. We discuss here shortly scoring functions designed for protein–protein docking, even though it might be extended to any kind of macromolecular docking (e.g., protein–peptide, protein–nucleotide, etc.). Due to the huge body of literature, we do not aspire to give the most comprehensive review covering all possible contributions, but strive to give an overview of various scoring functions and their features since those are also often used in the prediction of binding affinity changes upon mutation. We also discuss the time evolution of the different types of scoring functions and draw the parallel with the evolution of $\Delta\Delta G$ predictors (Figure 2).

The primal scoring function dates from late 1980s, when Wodak and Janin introduced the key components of any modern docking software.[74] Shape complementarity-based scoring method was implemented in this very first protein–protein docking simulation and since this original publication, it has been and still is widely used, often as one of the components of scoring functions in many modern docking programs.

Energy-based scoring functions have been the preferred alternative since the beginnings of the 21st century. They can be found in several different docking programs such as HADDOCK,[72,73] ZDOCK,[75] and PyDock.[76,77] Energy-based scoring functions are typically a weighted linear function of multiple energetic terms (e.g., electrostatic and van der Waals [intermolecular] energies, hydrogen bonding energy, empirical desolvation energy, etc.) and surface area-based terms, like the Buried Surface Area (BSA).

Like $\Delta\Delta G$ predictors, docking scoring functions have also seen the appearance of knowledge-based potentials (i.e., statistical potentials). Statistical potentials-based methods originated from Miyazawa and Jernigan's seminal work in 1985.[78] This line of methods estimates the interaction potentials of pairwise atom/residue types by converting their pair density observed from experimentally solved 3D structures into energies. The basic assumption is that the pairwise interaction of specific types of residues/atoms (a state) follows the Boltzmann distribution,[79–84] which elegantly describes the relationship between the probability of a state and its energy. A first implementation of a residue–residue statistical potential used for screening docking models can be found in the 3D-DOCK software suite.[85] Other representative examples of the use of statistical potentials in scoring are DFIRE[86] and SIPPER.[87]

A recent important advance in scoring functions comes from InterEvScore,[88] a method combining (co-)evolution with multibody statistical potentials. Similar to BindProfX which use conservation as a main feature for $\Delta\Delta G$ predictions, InterEvScore makes use of this information (when available) for scoring protein–protein docking models, outperforming several state-of-the-art methods in those cases. Clearly, the major drawback here is the availability of co-evolution data. Nevertheless with the drastic advances of sequencing technology and the improved performance of PPI partner predictions, we believe such limitation could be alleviated in near future.

Machine learning-based scoring functions[89–95] often treat the scoring problem as a binary classification problem, predicting a docked model as near-native or not. They usually integrate a large number of biophysical and sequence features (as those discussed previously) from the interface of the docked models. As a brief illustration of machine learning-based scoring functions, we selected two representative methods, IRaPPA[96] and SPIDER.[97]

IRaPPA integrates 91 physicochemical features, including physical and statistical potentials, empirical potentials, cluster size, etc. These features are combined to train ranking support vector machine (R-SVM) models that aim to minimize the fraction of swapped pairs relative to a perfect ranking. Based on the rankings from multiple R-SVM models, a consensus ranking is generated as final ranking through Schulze voting method.[98]

SPIDER is a structural binding motif-based method which uses a database of ~25,000 frequent subgraphs extracted from X-ray protein complexes. Those graphs represent the multiresidue interfacial motifs/patterns that appear frequently in native PPIs. The SPIDER score of a docked model consists of two main parts: (a) the frequency of the native binding motifs appearing at the interface of a docked model and (b) the geometrical fit between matching native patterns and the docked pattern.

4 | DATA! DATA! DATA!

Many $\Delta\Delta G$ predictors have been developed in the past decade, especially after the release of SKEMPI[28] in 2012. Although SKEMPI and other databases provide >7,000 experimental mutation data for ~240 protein–protein complexes, the amount of available data is still very limited compared to the number of biological complexes for which a 3D structure is available in the PDB (~5,500 to date) and especially to the estimated size of the interactome (e.g., >100,000 human PPIs[5]). Furthermore, both types and locations of the mutations are not evenly represented. It is therefore difficult to train $\Delta\Delta G$ predictors on these rather limited interactions and mutations datasets, and to obtain a consistent prediction accuracy on new interactions and mutations. This is well demonstrated by the large drop in performance of LOCOCV and on blind tests. These limitations make it difficult to take full advantage of machine learning methods, and in particular deep learning which is becoming increasingly popular.[99] Provided sufficient data, machine learning methods could improve the prediction performance because of their data-driven characteristic, while simple linear functions usually have limited performance improvement due to their simple form.[100] Currently, simple linear $\Delta\Delta G$ predictors still compete with machine learning-based $\Delta\Delta G$ predictors, indicating there is still a plenty of room to improve machine learning-based $\Delta\Delta G$ predictors. The limited number of experimental $\Delta\Delta G$ data not only limits the development of $\Delta\Delta G$ predictors but also their benchmarking and comparison since it is difficult to define truly blind test datasets that no single predictor has previously seen.

In terms of data generation, deep sequencing has the potential of generating much more thermodynamic information although the quality currently may be not very high. Further, the collection of mutation data, often a manual process based on mining literature, could greatly benefit from text recognition and natural language processing techniques to automatically fetch relevant data from literature. A similar approach has been recently published to extract information for the scoring of protein–protein complexes in docking.[101] It is also crucial to provide services and/or build an infrastructure to directly collect new data from the researchers generating them. The SKEMPI 2.0[32] website (https://life.bsc.es/pid/skempi2) provides such a service. Protabank[102] (https://www.protabank.org/) is another database, with a larger ambition, aiming at building a comprehensive databank for protein engineering data. All initiatives should ideally benefit from direct links with the PDB so that structure depositors are directly encouraged to deposit any thermodynamics and kinetics data they might have, both about the complexes themselves and also the impact of mutation on those.

5 | BLIND CHALLENGES AS CATALYSTS OF DEVELOPMENT

Blind challenges can help to foster the development of better $\Delta\Delta G$ predictors and scoring functions by evaluating methods in a real case scenario. CAPRI[103] is such a community-wide challenge for fully blind tests of computational docking algorithms. It involves two categories: (a) docking round, in which docking procedures are assessed for the prediction of biomolecular complexes and (b) scoring round, in which docking scoring functions are tested. From 2001, CAPRI has been playing a central role in stimulating the progress of docking algorithms and scoring functions as described in literature.[15,101–108]

In the 5th edition, CAPRI was extended to assess the methods to predict the impact of mutations for two designed influenza hemagglutinin binders based on yeast display enrichment data obtained using deep sequencing.[15,40,104] This assessment reported severe difficulties of docking scoring functions for predicting the impact of mutations on PPIs. This outcome should not be read as a failure but as a necessity of developing better $\Delta\Delta G$ predictors. Note that scoring and $\Delta\Delta G$ prediction are in principle different problems, and no $\Delta\Delta G$ predictors discussed above were assessed in this CAPRI challenge. Also, this was a one-time challenge in CAPRI because there has not been any other similar challenge since then. More such blind challenges are therefore highly needed in order to assess the state-of-the-art $\Delta\Delta G$ predictors and further catalyze developments in this field, as has been achieved by CAPRI for docking and scoring in PPIs and the D3R (Drug Design Data Resource) grand challenges[105,106] for protein–small molecule pose and binding affinity prediction.

6 | CONCLUSIONS AND OUTLOOK

We have reviewed here various aspects of the prediction of the impact of mutation on the binding affinity of protein–protein complexes, excluding the more rigorous free-energy perturbation methods. As for the scoring in protein–protein docking, the field has seen a transition from purely energetic functions to a combination with statistical potentials and increasing use of machine learning techniques. The quantity and quality of the available data, and their coverage in terms of type/variety of complexes, is still very much a limiting factor for training predictors that can demonstrate a good performance over a wide range of complexes. As an illustration of this, recent benchmarking on a large blind subset of the recently released SKEMPI 2.0 (487 mutations in 56 complexes), which none of the predictors had previously seen, showed that iSEE, FoldX, mCSM, and BindProfX[43] did not perform as well as on the training sets they originally used with PCC values all below 0.4. Even so, we have seen some promising contributions of such predictors to practical problems, for example, References 107–110.

To answer the original question in the title of this review, there is still plenty of work to be done to improve and further develop current predictors to make them reliable on fully blind and diverse datasets. Furthermore, they are also still very sensitive to the quality of the 3D structures used. With machine learning and in particular deep learning coming of age, there is a great potential to improve the field provided sufficient data become available. Here deep sequencing might contribute, with the catch that the quality of the generated data might be limited. Provided sufficient data, deep learning approaches might be able to deal with this problem. It should also be noted that several limiting factors identified for binding affinity prediction of PPIs[16] might also apply here, for example, ignoring the free state of the mutated protein and solvent effect. These might require special attention in the future. One should, however, not forget that some problems will still require more rigorous computational approaches such as free energy perturbation methods and related in order to obtain reliable predictions that also properly consider entropic effects.[111,112] These remain, however, computationally expensive and will, for the time being, not be applicable in high throughput manner for the thousands of mutations currently available in databases or to predict the impact of single nucleotide polymorphism on PPIs.[113] Community-wide blind challenges like CAPRI and D3R, and hopefully new ones, will remain important catalysts for the development of new methods and continuous improvement of existing ones.

ACKNOWLEDGMENTS

We thank Dr Adrien Melquiond (Utrecht University) for fruitful discussions. This study was supported by Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant no. 718.015.001). Netherlands eScience Center ASDI grant (No. 027016G04). European Union Horizon 2020 Program BioExcel grant (No. 675728). China Scholarship Council Fellowship (No. 201406220132).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLE

Interaction entropy for computational alanine scanning in protein–protein binding

ORCID

Cunliang Geng https://orcid.org/0000-0002-1409-8358
Li C. Xue https://orcid.org/0000-0002-2613-538X
Jorge Roel-Touris https://orcid.org/0000-0002-6588-624X
Alexandre M. J. J. Bonvin https://orcid.org/0000-0001-7369-1322

REFERENCES

1. Shoemaker BA, Panchenko AR. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol*. 2007;*3*:e42.
2. Melquiond ASJ, Karaca E, Kastritis PL, Bonvin AMJJ. Next challenges in protein–protein docking: From proteome to interactome and beyond. *WIREs Comput Mol Sci*. 2012;*2*:642–651.
3. Marsh JA, Teichmann SA. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem*. 2015;*84*:551–575. https://doi.org/10.1146/annurev-biochem-060614-034142
4. Chavez JD, Bruce JE. Chemical cross-linking with mass spectrometry: A tool for systems structural biology. *Curr Opin Chem Biol*. 2019;*48*:8–18.
5. Mosca R, Céol A, Aloy P. Interactome3D: Adding structural details to protein networks. *Nat Methods*. 2012;*10*:47–53.
6. Bai X-C, McMullan G, Scheres SHW. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci*. 2015;*40*:49–57.
7. Elmlund D, Le SN, Elmlund H. High-resolution cryo-EM: The nuts and bolts. *Curr Opin Struct Biol*. 2017;*46*:1–6.

8. Schmidt C, Urlaub H. Combining cryo-electron microscopy (cryo-EM) and cross-linking mass spectrometry (CX-MS) for structural elucidation of large protein assemblies. *Curr Opin Struct Biol*. 2017;*46*:157–168.
9. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*. 2002;*47*:409–443.
10. Huang S-Y. Search strategies and evaluation in protein–protein docking: Principles, advances and challenges. *Drug Discov Today*. 2014;*19*:1081–1096.
11. Huang S-Y. Exploring the potential of global protein–protein docking: An overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov Today*. 2015;*20*:969–977.
12. Rodrigues JPGLM, Bonvin AMJJ. Integrative computational modeling of protein interactions. *FEBS J*. 2014;*281*:1988–2003.
13. Soni N, Madhusudhan MS. Computational modeling of protein assemblies. *Curr Opin Struct Biol*. 2017;*44*:179–189.
14. Moal IH, Torchala M, Bates PA, Fernández-Recio J. The scoring of poses in protein–protein docking: Current capabilities and future directions. *BMC Bioinformatics*. 2013;*14*:286.
15. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. *Proteins*. 2013;*81*:2082–2095.
16. Kastritis PL, Bonvin AMJJ. On the binding affinity of macromolecular interactions: Daring to ask why proteins interact. *J Roy Soc Interface*. 2012;*10*:20120835.
17. Kastritis PL, Bonvin AM. Molecular origins of binding affinity: Seeking the Archimedean point. *Curr Opin Struct Biol*. 2013;*23*:868–877.
18. Vangone A, Oliva R, Cavallo L, Bonvin AMJJ. Prediction of biomolecular complexes. In: Rigden DJ, editor. *From protein structure to function with bioinformatics*. Dordrecht, The Netherlands: Springer, 2017; p. 265–292.
19. Gromiha MM, Yugandhar K, Jemimah S. Protein–protein interactions: Scoring schemes and binding affinity. *Curr Opin Struct Biol*. 2017;*44*:31–38.
20. Moreira IS, Fernandes PA, Ramos MJ. Hot spots—A review of the protein–protein interface determinant amino-acid residues. *Proteins*. 2007;*68*:803–812.
21. Fernández-Recio J. Prediction of protein binding sites and hot spots. *WIREs Comput Mol Sci*. 2011;*1*:680–698.
22. Chipot C. Frontiers in free-energy calculations of biological systems. *WIREs Comput Mol Sci*. 2014;*4*:71–89.
23. Steinbrecher T, Abel R, Clark A, Friesner R. Free energy perturbation calculations of the thermodynamics of protein side-chain mutations. *J Mol Biol*. 2017;*429*:923–929.
24. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. New York, NY: Springer Science & Business Media, 2013.
25. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statist Surv*. 2010;*4*:40–79.
26. Thorn KS, Bogan AA. ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*. 2001;*17*:284–285.
27. Kumar MDS. PINT: Protein–protein interactions thermodynamic database. *Nucleic Acids Res*. 2006;*34*:D195–D198.
28. Moal IH, Fernández-Recio J. SKEMPI: A structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*. 2012;*28*:2600–2607.
29. Geng C, Vangone A, Bonvin AMJJ. Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Eng Des Sel*. 2016;*29*:291–299.
30. Jemimah S, Yugandhar K, Michael Gromiha M. PROXiMATE: A database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*. 2017;*33*:2787–2788.
31. Liu Q, Chen P, Wang B, Li J. dbMPIKT: A web resource for the kinetic and thermodynamic database of mutant protein interactions; 2017.
32. Jankauskaite J, Jiménez-García B, Dapkūnas J, Fernández-Recio J, Moal IH. SKEMPI 2.0: An updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*. 2018;*9*:e1003216.
33. Sirin S, Apgar JR, Bennett EM, Keating AE. AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Sci*. 2015;*25*:393–409.
34. Borrman T, Cimons J, Cosiano M, et al. ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins*. 2017;*85*:908–916.
35. Dourado DFAR, Flores SC. Modeling and fitting protein–protein complexes to predict change of binding energy. *Sci Rep*. 2016;*6*:774.
36. Kowalsky CA, Whitehead TA. Determination of binding affinity upon mutation for type I dockerin-cohesin complexes from clostridium thermocellum and clostridium cellulolyticum using deep sequencing. *Proteins*. 2016;*84*:1914–1928.
37. Schreiber G, Fleishman SJ. Computational design of protein–protein interactions. *Curr Opin Struct Biol*. 2013;*23*:903–910.
38. Fowler DM, Fields S. Deep mutational scanning: A new style of protein science. *Nat Methods*. 2014;*11*:801–807.
39. Wrenbeck EE, Faber MS, Whitehead TA. Deep sequencing methods for protein engineering and design. *Curr Opin Struct Biol*. 2017;*45*:36–44.
40. Moretti R, Fleishman SJ, Agius R, et al. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. *Proteins*. 2013;*81*:1980–1987.
41. Harris DT, Wang N, Riley TP, et al. Deep mutational scans as a guide to engineering high affinity T cell receptor interactions with peptide-bound major histocompatibility complex. *J Biol Chem*. 2016;*291*:24566–24578.
42. Brender JR, Zhang Y. Predicting the effect of mutations on protein–protein binding interactions through structure-based interface profiles. *PLoS Comput Biol*. 2015;*11*:e1004494.
43. Xiong P, Zhang C, Zheng W, Zhang Y. BindProfX: Assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J Mol Biol*. 2017;*429*:426–434.
44. Berliner N, Teyra J, Çolak R, Garcia Lopez S, Kim PM. Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One*. 2014;*9*:e107353.
45. Riley TP, Ayres CM, Hellman LM, et al. A generalized framework for computational design and mutational scanning of T-cell receptor binding interfaces. *Protein Eng Des Sel*. 2016;*17*:87.
46. Rose PW, Prlić A, Altunkaya A, et al. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res*. 2017;*45*:D271–D281.
47. Pires DEV, Ascher DB, Blundell TL. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*. 2013;*30*:335–342.
48. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: Prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Res*. 2013;*41*:W333–W339.
49. Moal IH, Fernández-Recio J. Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation. *J Chem Theory Comput*. 2013;*9*:3715–3727.
50. Petukh M, Li M, Alexov E. Predicting binding free energy change caused by point mutations with knowledge-modified MM/PBSA method. *PLoS Comput Biol*. 2015;*11*:e1004276.
51. Li M, Simonetti FL, Goncearenco A, Panchenko AR. MutaBind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic Acids Res*. 2016;*44*:W494–W501.
52. Simões ICM, Costa IPD, Coimbra JTS, Ramos MJ, Fernandes PA. New parameters for higher accuracy in the computation of binding free energy differences upon alanine scanning mutagenesis on protein–protein interfaces. *J Chem Inf Model*. 2016;*57*:60–72.

53. Barlow KA, Ó Conchúir S, Thompson S, et al. Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *J Phys Chem B*. 2018;*122*:5389–5399.
54. Li M, Petukh M, Alexov E, Panchenko AR. Predicting the impact of missense mutations on protein–protein binding affinity. *J Chem Theory Comput*. 2014;*10*: 1770–1780.
55. Dourado DFAR, Flores SC. A multiscale approach to predicting affinity changes in protein–protein interfaces. *Proteins*. 2014;*82*:2681–2690.
56. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J Mol Biol*. 2002; *320*:369–387.
57. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: An online force field. *Nucleic Acids Res*. 2005;*33*:W382–W388.
58. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proc Natl Acad Sci U S A*. 2002;*99*:14116–14121.
59. Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. Predicting free energy changes using structural ensembles. *Nat Methods*. 2009;*6*:3–4.
60. Viricel C, de Givry S, Schiex T, Barbe S. Cost function network-based design of protein–protein interactions: Predicting changes in binding affinity. *Bioinformatics*. 2018;*34*:2581–2589.
61. Beard H, Cholleti A, Pearlman D, Sherman W, Loving KA. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein–protein complexes. *PLoS One*. 2013;*8*:e82849.
62. Petukh M, Dai L, Alexov E. SAAMBE: Webserver to predict the charge of binding free energy caused by amino acids mutations. *Int J Mol Sci*. 2016;*17*:547.
63. Pires DEV, Ascher DB. mCSM-AB: A web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res*. 2016;*44*:W469–W473.
64. Witvliet DK, Strokach A, Giraldo-Forero AF, Teyra J, Çolak R, Kim PM. ELASPIC web-server: Proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. *Bioinformatics*. 2016;*32*:1589–1591.
65. Geng C, Vangone A, Folkers GE, Xue LC, Bonvin AMJJ. iSEE: Interface structure, evolution and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins*. 2018.
66. Ackers GK, Smith FR. Effects of site-specific amino acid modification on protein interactions and biological function. *Annu Rev Biochem*. 1985;*54*:597–629.
67. Shapovalov MV, Dunbrack RL Jr. A smoothed backbone-dependent Rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*. 2011;*19*:844–858.
68. O'Meara MJ, Leaver-Fay A, Tyka MD, et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J Chem Theory Comput*. 2015;*11*:609–622.
69. Song Y, Tyka M, Leaver-Fay A, Thompson J, Baker D. Structure-guided forcefield optimization. *Proteins*. 2011;*79*:1898–1909.
70. Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol*. 2008;*380*:742–756.
71. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009;*4*: 1073–1081.
72. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A protein−protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*. 2003;*125*:1731–1737.
73. van Zundert GCP, Rodrigues JPGLM, Trellet M, et al. The HADDOCK2.2 web server: User-friendly integrative modeling of biomolecular complexes. *J Mol Biol*. 2016;*428*:720–725.
74. Wodak SJ, Janin J. Computer analysis of protein–protein interaction. *J Mol Biol*. 1978;*124*:323–342.
75. Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z. ZDOCK server: Interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics*. 2014;*30*:1771–1773.
76. Cheng TMK, Blundell TL, Fernández-Recio J. pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins*. 2007; *68*:503–515.
77. Jiménez-García B, Pons C, Fernández-Recio J. pyDockWEB: A web server for rigid-body protein–protein docking using electrostatics and desolvation scoring. *Bioinformatics*. 2013;*29*:1698–1699.
78. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*. 1985;*18*:534–552.
79. Huang S-Y, Zou X. An iterative knowledge-based scoring function for protein–protein recognition. *Proteins*. 2008;*72*:557–579.
80. Huang S-Y, Zou X. An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. *J Comput Chem*. 2006;*27*:1866–1875.
81. Huang S-Y, Zou X. An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring function. *J Comput Chem*. 2006;*27*:1876–1882.
82. Huang S-Y, Zou X. MDockPP: A hierarchical approach for protein–protein docking and its application to CAPRI rounds 15–19. *Proteins*. 2010;*78*:3096–3103.
83. Liu S, Vakser IA. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein–protein docking. *BMC Bioinformatics*. 2011;*12*:280.
84. Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein−ligand, protein−protein, and protein−DNA complexes. *J Med Chem*. 2005;*48*: 2325–2335.
85. Moont G, Gabb HA, Sternberg MJE. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*. 1999;*35*:364–373.
86. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*. 2002;*11*:2714–2726.
87. Pons C, Talavera D, la Cruz de X, Orozco M, Fernández-Recio J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): A new efficient potential for protein−protein docking. *J Chem Inf Model*. 2011;*51*:370–377.
88. Andreani J, Faure G, Guerois R. InterEvScore: A novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics*. 2013;*29*:1742–1749.
89. Palma PN, Krippahl L, Wampler JE, Moura JJG. BiGGER: A new (soft) docking algorithm for predicting protein interactions. *Proteins*. 2000;*39*:372–384.
90. Bordner AJ, Gorin AA. Protein docking using surface matching and supervised machine learning. *Proteins*. 2007;*68*:488–502.
91. Bernauer J, Azé J, Janin J, Poupon A. A new protein–protein docking scoring function based on interface residue properties. *Bioinformatics*. 2007;*23*:555–562.
92. London N, Schueler-Furman O. Funnel hunting in a rough terrain: Learning and discriminating native energy funnels. *Structure*. 2008;*16*:269–279.
93. Chae MH, Krull F, Lorenzen S, Knapp EW. Predicting protein complex geometries with a neural network. *Proteins*. 2009;*78*:1026–1039.
94. Bourquard T, Bernauer J, Azé J, Poupon A. A collaborative filtering approach for protein–protein docking scoring functions. *PLoS One*. 2011;*6*:e18541.
95. Fink F, Hochrein J, Wolowski V, Merkl R, Gronwald W. PROCOS: Computational analysis of protein–protein complexes. *J Comput Chem*. 2011;*32*:2575–2586.
96. Moal IH, Barradas-Bautista D, Jiménez-García B, et al. IRaPPA: Information retrieval based integration of biophysical models for protein assembly selection. *Bioinformatics*. 2017;*33*:1806–1813.

97. Khashan R, Zheng W, Tropsha A. Scoring protein interaction decoys using exposed residues (SPIDER): A novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins*. 2012;*80*:2207–2217.

98. Schulze M. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Soc Choice Welf*. 2010;*36*: 267–303.

99. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol*. 2016;*12*:878.

100. Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *WIREs Comput Mol Sci*. 2015;*5*:405–424.

101. Badal VD, Kundrotas PJ, Vakser IA. Natural language processing in text mining for structural modeling of protein complexes. *BMC Bioinformatics*. 2018;*19*:84.

102. Wang CY, Chang PM, Ary ML, et al. ProtaBank: A repository for protein design and engineering data. *Protein Sci*. 2018;*27*:1113–1124.

103. Janin J. Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins*. 2002;*47*:257–257.

104. Yin J, Henriksen NM, Slochower DR, et al. Overview of the SAMPL5 host–guest challenge: Are we doing better? *J Comput Aided Mol Des*. 2016;*31*:1–19.

105. Gathiaka S, Liu S, Chiu M, et al. D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J Comput Aided Mol Des*. 2016;*30*: 651–668.

106. Gaieb Z, Liu S, Gathiaka S, et al. D3R grand challenge 2: Blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des*. 2017;*32*:1–20.

107. Usher JL, Ascher DB, Pires DEV, Milan AM, Blundell TL, Ranganath LR. Analysis of HGD gene mutations in patients with Alkaptonuria from the United Kingdom: Identification of novel mutations. *JIMD reports*. Volume 24. Berlin, Heidelberg: Springer, 2014; p. 3–11.

108. Studer RA, Rodriguez-Mias RA, Haas KM, et al. Evolution of protein phosphorylation across 18 fungal species. *Science*. 2016;*354*:229–232.

109. Nosrati M, Solbak S, Nordesjö O, et al. Insights from engineering the Affibody-fc interaction with a computational-experimental method. *Protein Eng Des Sel*. 2017;*30*:593–601.

110. Khor BY, Lim TS, Noordin R, Choong YS. The design of target specific antibodies (scFv) by applying de novo workflow: Case study on BmR1 antigen from Brugia malayi. *J Mol Graph Model*. 2017;*76*:543–550.

111. Kollman PA, Massova I, Reyes C, et al. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc Chem Res*. 2000;*33*:889–897.

112. Mobley DL, Gilson MK. Predicting binding free energies: Frontiers and benchmarks. *Annu Rev Biophys*. 2017;*46*:531–558.

113. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montaño B, Blundell TL, Ascher DB. Mutations at protein–protein interfaces: Small changes over big surfaces have large impacts on human health. *Prog Biophys Mol Biol*. 2016;*128*:3–13.

How to cite this article: Geng C, Xue LC, Roel-Touris J, Bonvin AMJJ. Finding the $\Delta\Delta G$ spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *WIREs Comput Mol Sci*. 2019;9: e1410. https://doi.org/10.1002/wcms.1410