Bioinformatics, 2018, 1–8
doi: 10.1093/bioinformatics/bty635
Advance Access Publication Date: 18 July 2018
Original Paper

Structural bioinformatics

SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation

Justina Jankauskaitė[1], Brian Jiménez-García[2,3], Justas Dapkūnas[1], Juan Fernández-Recio[2,4] and Iain H. Moal[5,*]

[1]Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius LT-10257, Lithuania, [2]Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain, [3]Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, the Netherlands, [4]Institut de Biologia Molecular de Barcelona (IBMB), CSIC, Barcelona 08028, Spain and [5]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, UK

*To whom correspondence should be addressed.

Associate Editor: Ioannis Xenarios

Received on May 16, 2018; revised on May 16, 2018; editorial decision on July 11, 2018; accepted on July 17, 2018

Abstract

Motivation: Understanding the relationship between the sequence, structure, binding energy, binding kinetics and binding thermodynamics of protein–protein interactions is crucial to understanding cellular signaling, the assembly and regulation of molecular complexes, the mechanisms through which mutations lead to disease, and protein engineering.

Results: We present SKEMPI 2.0, a major update to our database of binding free energy changes upon mutation for structurally resolved protein–protein interactions. This version now contains manually curated binding data for 7085 mutations, an increase of 133%, including changes in kinetics for 1844 mutations, enthalpy and entropy changes for 443 mutations, and 440 mutations, which abolish detectable binding.

Availability and implementation: The database is available as supplementary data and at https://life.bsc.es/pid/skempi2/.

Contact: moal@ebi.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein–protein interactions are central to almost all biological processes, from cellular signal transduction and the assembly of mesoscopic structures such as myofilaments, to viral adhesion and the immune response. Consequently the effects of changes in protein sequence on the structure, thermodynamics and kinetics of protein–protein interactions has wide implications for constraining the permissible substitutions that accrue over the course of evolution, and for understanding the molecular etiology of disease. Methods which measure, predict or optimize these changes have applications in designing *de novo* interactions (Fleishman *et al.*, 2011), enhancing the specificity and affinity of biological therapeutics (e.g. Arkadash *et al.*, 2017), designing combinatorial protein libraries (e.g. Guntas *et al.*, 2010), uncovering the effects of pathological mutations (e.g. Tidow *et al.*, 2006), locating druggable binding sites (e.g. Stevers *et al.*, 2017) and binding hotspots for drug design (Guo *et al.*, 2014), altering binding kinetics (Cohen-Khait and Schreiber, 2016; Rosenfeld *et al.*, 2017), protein–protein docking (e.g. Epa *et al.*, 2013), and characterizing transition states (e.g. Wu *et al.*, 2002), binding pathways (Plattner *et al.*, 2017), and sequence-affinity landscapes (Aizner *et al.*, 2014).

SKEMPI is a manually curated database of mutations in structurally characterized protein–protein interactions and the effect of those mutations on binding affinity and other parameters (Moal and

Fernandez-Recio, 2012). The first release has been used as a basis for many further studies, including the development of energy functions (Moal and Fernandez-Recio, 2013; Moal *et al.*, 2015b) which were subsequently implemented in the CCharPPI web server for characterizing protein–protein interactions (Moal *et al.*, 2015a), as well as being used for ranking docked poses (Barradas-Bautista *et al.*, 2017; Moal *et al.*, 2013; Moal *et al.*, 2017; Pfeiffenberger *et al.*, 2017). SKEMPI has also been used to study human disease (Das *et al.*, 2014; Peng and Alexov, 2016; Petukh *et al.*, 2015a), assessing the role of dynamics on binding (Sumbul *et al.*, 2015), exploring the conservation of binding regions (Hu *et al.*, 2016), evaluating experimental affinity measurement methods (Geng *et al.*, 2016), as well serving as a data source for models which predict dissociation rate changes upon mutation (Agius *et al.*, 2013), pathological mutations (Gossage *et al.*, 2014), hotspot residues (e.g. Hwang *et al.*, 2014; Liu *et al.*, 2015; Melo *et al.*, 2016; Simoes *et al.*, 2017) and changes in binding energy (e.g. Barlow *et al.*, 2018; Berliner *et al.*, 2014; Dehouck *et al.*, 2013; Dourado and Flores, 2014; Lai *et al.*, 2017; Li *et al.*, 2016; Moretti *et al.*, 2013; Pallara *et al.*, 2013; Pantazes *et al.*, 2015; Petukh *et al.*, 2015b; Pires *et al.*, 2014; Xiong *et al.*, 2017; Yan *et al.*, 2017; Zhao *et al.*, 2014).

Here we present a major update to the benchmark in terms of the number of mutations in the database and the number of different systems included (Table 1). We now also include details of the experimental method for all entries, based on the categories of Geng *et al.*, 2016, as well as mutations which abolished detectable binding or for which only an upper or lower affinity limit could be ascertained for the wild-type or mutant.

2 Materials and methods

2.1 Data sources

Just over two fifths of the data come from the previous version of SKEMPI (Moal and Fernandez-Recio, 2012), comprised mostly of data found in literature sources which came to the authors' attention, in some cases during the data collection for the structural affinity benchmark (Kastritis *et al.*, 2011) and following references therein. Some entries in SKEMPI 1.1 were found by checking the references in the ASEdb (Thorn and Bogan, 2001) and PINT (Kumar and Gromiha, 2006) databases, although not all the data passed the checks required for inclusion (see Section 2.2). Similarly, most of the new data in SKEMPI 2.0 was found by searching the literature, partly in tandem with the literature search for the more recent structural affinity benchmark (Vreven *et al.*, 2015). During data collection three other relevant databases were published: ABbind (Sirin *et al.*, 2016), PROXiMATE (Jemimah *et al.*, 2017) and dbMPIKT (Liu *et al.*, 2017). Their references were checked if they were not already included. Data from these sources comprise 4%, 3% and 6% of SKEMPI 2.0 respectively. As with ASEdb and PINT, none of the data were directly copied into SKEMPI. Moreover, the cited papers were read and data entered using the same checks and procedures as other entries.

Table 1. Comparison with previous version

	SKEMPI 1.1	SKEMPI 2.0
Entries	3047	7085
Unique entries	2792	6187
k_{on} and k_{off}	713	1844
ΔH and ΔS	127	443
Inequalities/no binding	0	440
Number of interactions	87	237
Number of PDB entries	158	345
Number of papers cited	66	295

2.2 Data collection

Each entry was found in the literature and manually vetted. To ensure quality, a number of stringent checks were applied. Firstly, we ensured that the structure and the paper reporting the affinities refer to the same protein in the same species, and that structural and affinity data matched in terms of cofactors, ancillary chains and post-translational modifications. For instance, we distinguish between RasGTP and RasGppNHp, as the nucleotide modulates the affinity of Ras with its effectors. Where the full-length protein was not used, we checked to ensure that the fragment in the crystal structure matched that for which affinities are reported.

Once the checks are passed, the data is collected, including the PDB file, the chains of the interacting subunits, the mutation, the wild-type and mutant affinities (K_D, M), the reference, the names of the proteins, the temperature at which the experiment is performed (T, K), the experimental method used (an extension of the category scheme of Geng *et al.*, 2016), notes on the entry and, when available, the association rate (k_{on}, $M^{-1}s^{-1}$), dissociation rate (k_{off}, s^{-1}), enthalpy (ΔH, kcal.mol^{-1}) and entropy (ΔS, cal.mol^{-1}.K^{-1}). For cases where multiple PDB entries are available, the higher resolution structure is chosen. Where affinities or kinetic or thermodynamic parameters are reported in different units, these are converted to the units specified above. In some cases, when not reported directly, K_D, k_{on}, k_{off}, ΔH and ΔS were calculated using the relationships $\Delta G = \Delta H - T\Delta S = RT\ln(K_D)$ and $K_D = 1/K_A = k_{off}/k_{on}$. To ensure consistency, the residue numbering in SKEMPI is the same as that reported in the PDB file. Thus, the numbering is often shifted or altered compared to that in the cited paper such that, for instance, if a crystal structure of an antibody is reported in the Kabat numbering scheme but the mutation data is not, then the mutation data is converted before entry into SKEMPI. For all entries it is the case that the affinity reported in the "wild-type" column corresponds to that of the PDB file, and the affinity in the "mutant" column is that after applying the specified mutation to the protein in the PDB file. Thus, where there are cases in which the PDB reports a mutant form and the entry corresponds to the reverse mutation back to the wild-type, the affinity of the former appears in the "wild-type" column and the latter in the "mutant" column. Such cases are noted in the database.

In addition to checking new entries, we reappraised the papers cited in SKEMPI 1.1 to collect data that were not collected previously, specifically to find mutants which abolish binding and to classify the experimental method used when not already included in the subset of SKEMPI covered in Geng *et al.*, 2016. It is worth noting that often an author's decision to report an interaction of affinity below the detection threshold as either non-binding, or as less affine than the weakest affinity presented in the paper, is arbitrary. Thus, those wishing to use the non-binding data as an inequality on the affinity may do so. We also corrected entries for five wild-type and four mutant affinities identified by Geng *et al.*, 2016.

2.3 Post-processing and annotation

In addition to the above data, SKEMPI 2.0 also provides data on the location of the mutated residues, the homology between interactions in the dataset, and processed PDB files, which can be easily parsed.

Residue location: Each mutated residue is classified according to the scheme proposed by Levy (2010); residues at the interface are classified as support (mostly buried when unbound and entirely

buried upon binding), core (mostly solvent exposed when unbound but buried upon binding) and rim (partly buried upon binding), while residues away from the binding site are classified as interior or surface. Solvent exposed surface area was calculated using CCP4 (Winn *et al.*, 2011).

Processed PDB files: The PDB files for the interactions, as downloaded from the Protein Data Bank (Berman *et al.*, 2000), often contain multiple copies of the interacting proteins in the unit cell or other chains irrelevant to the interaction. In one instance, the binding of dimeric myostatin to follistatin-like 3, the myostatin dimer must be created by tessellating the unit cell. Further, some PDB files contain features that are not readily parsed by some software, such as residue insertion codes or negative residue numbers. To help users we provide "cleaned" PDB files which contain only the chains of interest, renumbered from one, as well as waters and other molecules with a non-hydrogen atom within 5 Å of a non-hydrogen atom of any of the chains of interest. Consequently, each mutation is reported with both PDB numbering and renumbered.

Defining homologous interactions: Each entry also specifies which other entries are mutations to homologous interactions. Two interactions are deemed homologous if they have a shared binding partner or homologous binding partner and at least 70% of the corresponding interface residues are common to both interactions. We determine the homology between proteins using the GAP4 program (Huang and Brutlag, 2007), and define homologous proteins as those with a similarity score greater than 50 and at least 30% sequence identity. Interface residues are defined as those with a non-hydrogen atom within 10 Å of a non-hydrogen atom on the binding partner. Interactions falling within manually assigned clusters of homologous interactions are designated as pMHC/TCR, antibody/antigen or protease/inhibitor. While the names of these clusters have been chosen to reflect the predominant function of their constituent interactions, they reflect the homologies within the dataset and are not functional assignments. Thus, for instance, some nanobodies are classified as antibodies as they bind to the same site as cetuximab, 14.3.d is classified as TCR, even though it is only the β chain, and its binding partner, enterotoxin C3, is classified as a pMHC.

3 Results and discussion

3.1 Diversity, bias and interrelationships within SKEMPI

In total 7085 entries were collected, summarized in Table 1 and Figure 1A. These data were derived from the literature and consequently, while encompassing a broad range of residues, proteins, interactions and systems, are biased toward the interests and capabilities of the research community. These biases are evident in the composition of the database according to parameters shown in Figure 1. The $\Delta\Delta G$ values span a large range, but mostly fall within –3 to 7 kcal.mol^{-1} (Fig. 1B), for both biophysical and technical reasons (see Section 3.3). Almost three quarters of the data correspond to single point mutations, and more than half of those are mutations to alanine (Fig. 1C). Charge swap mutations and mutations between aromatic residues are also over-represented. Most single point mutations are located at the binding site, and most of those are at the core of the interface. Similarly, most double mutations are both in the binding site, and most of those are both in the core. By far the most popular methods for measuring binding affinity were surface plasmon resonance and spectroscopic methods such as fluorescence. Large biases toward specific interactions and classes of interaction are also present, such as early studies into protease inhibition and immunological interactions such as antibody-antigen complexes, the recognition of peptides presented on cell surfaces by T-cell receptors, cytokine signaling and the complement system. Indeed, almost half of the data corresponds to protease-inhibitor, antibody-antigen and pMHC/TCR interactions alone. While many interactions within these classes share common binding sites or homologous binding sites (Fig. 2), there are also connections between these groups, for instance via inhibitory antibodies which bind to a protease active site, or due to common binding regions of proteins in the immunoglobulin superfamily, such as antibodies, TCRs, MHCs and β-2 microglobulin. In addition, present in the data are smaller clusters of shared and homologous interactions, such as the Ras-effector cluster. These relationships are noted in the database and may be useful for avoiding overfitting when developing models or for validation and estimating generalization error, as described previously (Moal and Fernandez-Recio, 2012).

The entries also vary in the degree of structural order. While most correspond to interactions between folded domains, the database contains entries in which structuring occurs upon binding, such as protein–peptide interactions and, in the extreme case, the ACTR/NCBD interaction in which both binding partners become ordered upon binding (Jemth *et al.*, 2014). Indeed, the requirements of having a structure in order to be included in the dataset, *ipso facto* biases the data, and means that there is no representation of "fuzzy" complexes in which a diffuse structural ensemble in the bound state prevents the formation of a resolvable crystal.

Another source of variation is the origins of the interacting proteins. While entries range from viral and bacterial to the higher eukaryotes, biases are evident in the over-representation of model organisms including humans. With the exception of the pMHC-TCR, antibody-antigen and protease-inhibitor classes, most of the interactions are endogenous. Nevertheless, the set also includes exogenous interactions ranging from those between proteins from different individuals within the same species, namely the sex fusion proteins Juno and Izumo1 from human sperm and egg respectively (Aydin *et al.*, 2016), to host-pathogen interactions such as adenovirus and coronavirus interactions with human receptors during viral entry (Howitt *et al.*, 2003; Seiradake *et al.*, 2006), to the inhibition of acetylcholinesterase by the snake venom neurotoxin fasciculin (Aizner *et al.*, 2014). For the pMHC-TCR interactions, there are a variety of presented antigens, including exogenous viral peptides and gluten, as well as endogenous autoimmune and cancer peptides. The antibody interactions include pathogen antibodies, as well as antibodies raised and optimized to target extracellular therapeutic targets. The protease interactions are mostly exogenous, arising from their inherent cross-reactivity due to the convergent evolution of their canonical inhibitory loop.

3.2 Notable studies comprising SKEMPI 2.0

The investigations from which SKEMPI data is derived are diverse, spanning many biological processes and reported in 295 publications including systematic scans, alanine and homolog scanning, design studies including computational design and designs derived from phage display, double mutant cycle studies, antibody engineering, biologic drug design and the evaluation of pathological mutations. The largest contribution comes from the group of the late Michael Laskowski Jr., a systematic study of all possible mutations at selected sites in the turkey ovomucoid third domain and its inhibitory interactions with four proteases (Lu *et al.*, 2001), as well as studies of interactions of the same domain in other bird species and the design of ultra-high affinity broad-spectrum inhibitors. Substantial data also come from investigations into the inhibitory

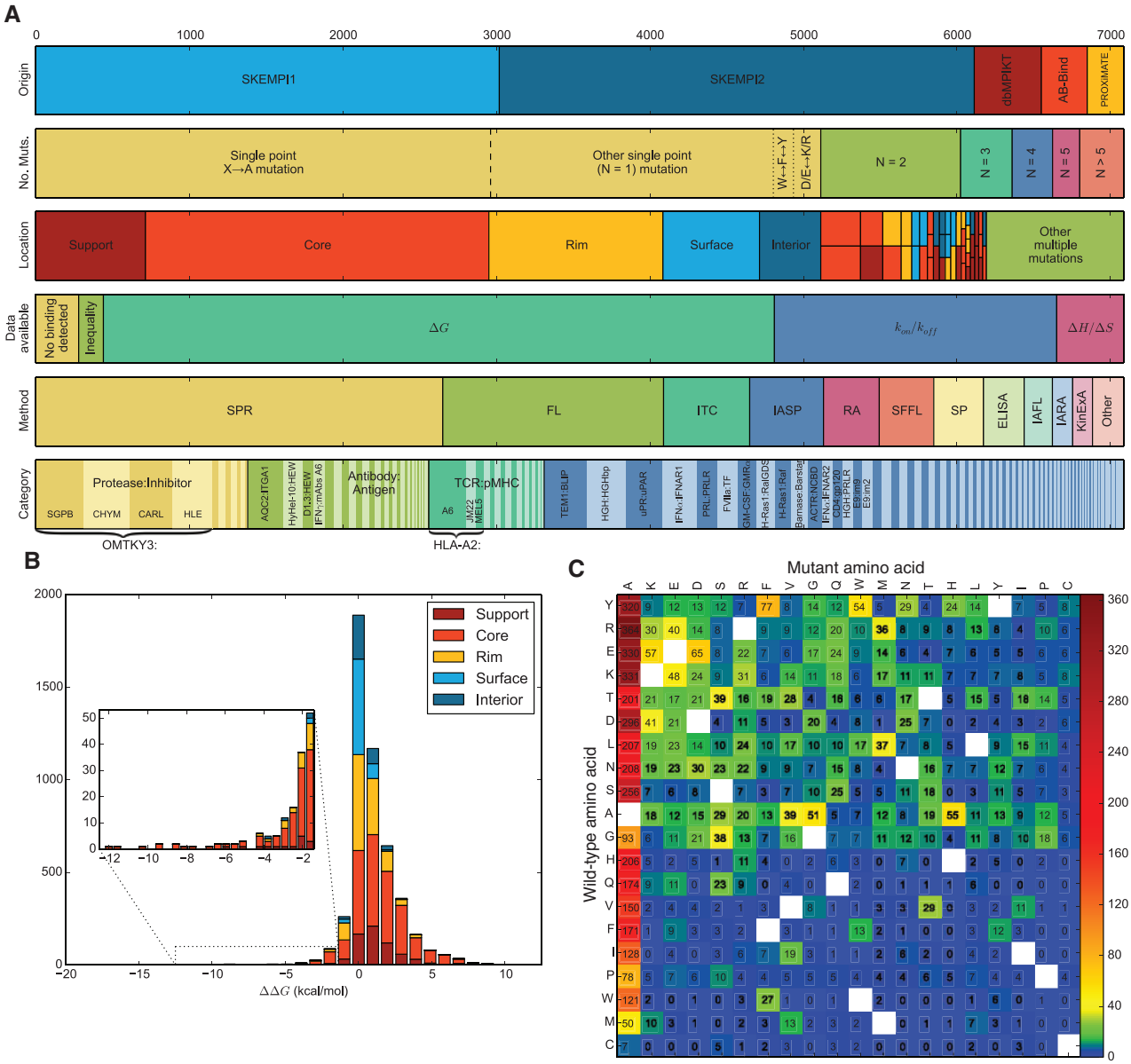

Fig. 1. An overview of SKEMPI 2.0. (**A**) Mutations partitioned according to their origin, the number of altered residues, location within the complex, by the availability of additional kinetic and thermodynamic data, according to the experimental method used, and by category. (**B**) Distribution of $\Delta\Delta G$. (**C**) Source and target amino acids for single point mutations

interactions of class A β-lactamases from the groups of Gideon Schreiber (e.g. Reichmann *et al.*, 2007) and Timothy Palzkill (e.g. Brown *et al.*, 2013), as well as cytokine receptor interactions, in particular studies of type I interferons also from the Schreiber group (e.g. Roisman *et al.*, 2001) and that of K. Christopher Garcia (Thomas *et al.*, 2011), but also the study of the GM-CSF/GMRα interaction from the group of Michael W. Parker (Broughton *et al.*, 2016). Other prominent sources of data are studies into hormone receptor interactions, in particular the human growth hormone receptor from the group of Jim Wells (e.g. Cunningham and Wells, 1991) and the prolactin receptor from the group of Michael E. Hodsdon (Kulkarni *et al.*, 2010), as well as studies into antigen recognition including the combined computational and experimental design study to enhance affinity of the AQC2 antibody to integrin α-1 from the group of Herman Van Vlijmen (Clark *et al.*, 2006), the dissection of the interactions of broadly neutralizing antibodies targeting HIV gp120 (Clark *et al.*, 2017) from the group of Richard A. Friesner, and various investigations from the group of Roy A. Mariuzza (e.g. Dall'Acqua *et al.*, 1998). Also notable is the alanine scanning of the urokinase-type plasminogen activator and its receptor from the group of Michael Ploug (Gardsvoll *et al.*, 2006), studies of Ras effector interactions from the group of Christian Herrmann (e.g. Kiel *et al.*, 2004), investigations of pMHC/TCR interactions from the group of Brian Baker (e.g. Piepenbrink *et al.*, 2013), and investigations into the cognate and non-cognate recognition of *Escherichia coli* colicin DNase bacteriotoxins by their immunity proteins from the group of Colin Kleanthous (e.g. Li *et al.*, 1997).

3.3 Range and error

Range: The changes in binding free energy upon mutation range from –12.4 to 12.4 kcal.mol^{-1}, as in SKEMPI 1.1, with $\Delta \log_{10} k_{on}$ ranging from –3.6 to 2.4, $\Delta \log_{10} k_{off}$ ranging from –6.0 to 6.8, $\Delta\Delta H$

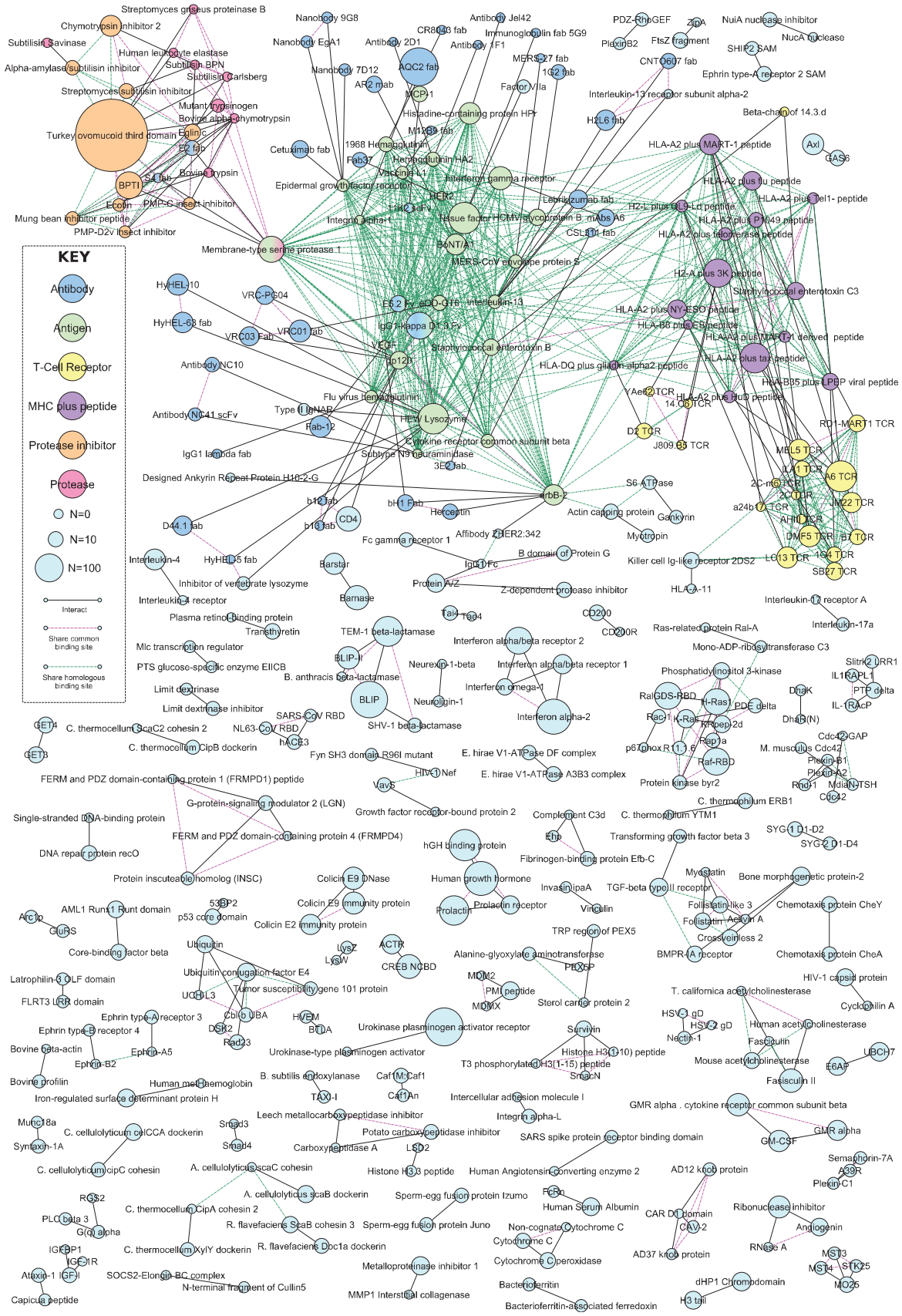

Fig. 2. Overview of the interactions in SKEMPI. Nodes indicate proteins, scaled by the number of mutations of that protein and coloured according to category. Edges show direct interactions, as well as relationships between proteins that share a common or homologous binding site

ranging from –18.3 to 26.5 kcal.mol^{-1}, and $\Delta\Delta S$ ranging from –61 to 80 cal.mol^{-1}.K^{-1}. Around 60 mutants are very destabilizing, reducing binding energy by 8 kcal.mol^{-1} or more. These are all in enzyme/inhibitor complexes such as the inhibition of acetylcholinesterase by the snake venom fasciculin, or the inhibition of enzymes which would be detrimental should they unbind and become active in the wrong location, such as nucleases (barnase/barstar, colicin E9 DNase/Im9, RNase A/angiogenin) and proteases (such as trypsin/BPTI). These interactions tend to be around picomolar affinity and are at the upper limit of what can be detected, due to the time required to reach equilibrium and the low concentrations required by the mass action law to probe informative regions of the binding curve. These very destabilizing mutations reduce affinity into the micromolar range, near the lower limit of what can be quantified using standard methods. As a consequence of both mutant and wild-type affinities being near detection thresholds, errors in these entries are typically large. Further, while some mutations may cause changes in affinity larger than seven orders of magnitude, the absence of affinities for such mutations in the benchmark can be explained by the fact that such mutations would involve affinities beyond the upper or lower limit. Indeed, there are new entries in which single or double substitutions reduce binding from tens of picomolar to having no detectable binding. For many of the highly destabilizing mutations a crystal structure for the mutant has also been solved, and the 30 most stabilizing mutations in the database ($\Delta\Delta G$ ¡ –5 kcal.mol^{-1}) consist of the reverse mutation applied to these structures. These are mostly single or double mutants, but include mutations to up to 27 residues of the non-cognate Colicin E2/Im9 complex, which move it toward the cognate E9/Im9 in sequence space (Li *et al.*, 1997).

Errors: Standard errors in K_D are typically reported in the order of 50%, around 0.25 kcal.mol^{-1}. These estimates are derived by repeat measurements using the same equipment, environment and protocol, and thus do not include errors arising from systematic bias. Such biases can, however, be estimated from pairs of entries in which the same mutation is evaluated by different groups or using different techniques. For 84% of 1741 such pairs, both entries give a $\Delta\Delta G$ value within 1 kcal.mol^{-1} of each other. For 704 pairs for which k_{on} is available for both, 80% have $\Delta \log_{10}(k_{on})$ within 0.5 of each other. For 702 pairs for which k_{off} is available, 83% have $\Delta \log_{10}(k_{off})$ within 0.5. For the 62 pair with both $\Delta\Delta H$ and $\Delta\Delta S$ values, 61% have $\Delta\Delta H$ within 3.0 kcal.mol^{-1} of each other and 58% have $\Delta\Delta S$ within 10 cal.mol^{-1}.K^{-1} of each other.

3.4 Mutant cycles

Within SKEMPI, some entries can be combined to construct mutant cycles, which quantify the interactions between residues, the dependence of these interactions on other residues, and other higher order effects. The most common instances are double mutant cycles, where affinities are available for the wild type, A, B and AB mutations, of which there are 610 examples. Of these, 53 involve at least one mutant for which binding was not observed, or only an inequality is available, and 235 involve mutations reported in the same reference, and thus the affinities are likely to have been measured using the same technique and conditions. A further 218 double mutant cycles can be constructed in the background of a third mutation (i.e. C, AC, BC and ABC mutations are available), of which 209 are not composed of non-binding mutations or mutations with inequalities, and 131 involve affinities coming from the same reference. Of the 766 double mutant cycles containing neither inequalities nor non-binding mutants, a number of parameters can be calculated, including $\Delta\Delta G_{ab\rightarrow Ab}$, $\Delta\Delta G_{ab\rightarrow aB}$ and $\Delta\Delta G_{ab\rightarrow AB}$, the binding free energy change of both single and the double mutation respectively, as well as $\Delta\Delta G_{aB\rightarrow AB}$ and $\Delta\Delta G_{Ab\rightarrow AB}$, the energy of a single mutation within the context of the other mutation, and $\Delta\Delta G_{int} = \Delta\Delta G_{aB\rightarrow AB} - \Delta\Delta G_{ab\rightarrow Ab} = \Delta\Delta G_{Ab\rightarrow AB} - \Delta\Delta G_{ab\rightarrow aB}$, the interaction energy of the two mutations (Horovitz and Fersht, 1990). From these, it can be deduced that 345 are additive ($\Delta\Delta G_{int} < 0.5$kcal.mol^{-1}). Of the non-additive cycles, 293 exhibit tighter binding in the double mutant than the sum of the single mutants (positive epistasis), of which six result in even tighter binding than individual effects of two single mutations that strengthen the interaction (synergistic positive), while 273 correspond to double mutants which reduce binding by less than the sum of two single mutants which reduce binding (antagonistic positive). Similarly, 128 cycles have double mutants exhibiting weaker binding than the sum of the two single mutants (negative epistasis), of which 58 contain two destabilizing single mutations (synergistic negative) and 26 contain two stabilizing single mutations (antagonistic negative). The range of $\Delta\Delta G_{int}$ values rarely fall outside of the –5 to 3 kcal.mol^{-1} range. Of the 421 non-additive cycles, 151 show noticeable sign epistasis, in which the sign of the effect of either the A or B mutation flips depending on the presence or absence of the background mutation (i.e. for the A mutation, $|\Delta\Delta G_{ab\rightarrow Ab}| > 0.2$ kcal.mol^{-1} and $|\Delta\Delta G_{aB\rightarrow AB}| > 0.2$ kcal.mol^{-1} and $|\Delta\Delta G_{ab\rightarrow Ab} - \Delta\Delta G_{aB\rightarrow AB}| > 0.4$ kcal.mol^{-1}). Of these, 38 correspond to mutations, which destabilize the complex in the presence of the background mutation, but stabilize in its absence (destabilizing sign epistasis), while 113 correspond to mutations, which stabilize the complex in the mutant background but otherwise destabilize the complex (stabilizing sign epistasis). Only eight cycles exhibit the more extreme reciprocal sign epistasis, which in six cases are where both single mutations are stabilizing (< -0.2 kcal.mol^{-1}), but the double mutant is destabilizing (>0.2 kcal.mol^{-1}), and the remaining two correspond to two destabilizing mutations (>0.2 kcal.mol^{-1}) for which the double mutation is stabilizing (< -0.2 kcal.mol^{-1}). The types of substitutions that can give rise to extreme effects such as stabilizing reciprocal sign epistasis can be illustrated with the MlcIIBGlc interaction in *E. coli* (Nam *et al.*, 2008), in which the removal of the F136 side-chain of MlC creates a large cavity at the binding interface, the addition of a phenylalanine at the A451 position of IIBGlc creates a large clash, however the double mutation creates an interaction that is even more stable than the wild-type by creating an anchor residue across the binding interface in which the cavity in MlC is filled by the new side-chain of IIB.

Higher order interaction terms can be garnered from higher cycles, such as triple mutant cubes, constructed from the energies of the wild-type, three single mutants, three corresponding double mutants and the triple mutant (Horovitz and Fersht, 1990). In SKEMPI, 45 triple mutant cubes can be made, with 10 coming from the same reference. For these, third order interaction energies fall within the –1 to 1 kcal.mol^{-1} range. For fourth order interactions, constructed from energies of the wild-type, four single mutants, six double mutants, four triple mutants and the quadruple mutant, 14 examples exist within SKEMPI. However, care should be taken in ascribing meaning to fourth order residue coupling energies due to the accumulations of errors, which in these cases are exacerbated by the affinities having been reported in different publications. No fifth or higher order interactions are present.

3.5 The SKEMPI website

The database is accessible online at https://life.bsc.es/pid/skempi2/, where the raw CSV (comma-separated values) file containing all the

data can be downloaded. The data can also be browsed online, ordered and searched by any field, such as the experimental method or the location of the mutation, or searching for a specific protein by its name or PDB code, and structures may be visualized. Other pages on the web site offer a summary of the data, an FAQ and help page, and a page for user contributions, which will be evaluated to appear in future releases.

Funding

This work has been supported by the European Molecular Biology Laboratory [I.H.M.]; Biotechnology and Biological Sciences Research Council [Future Leader Fellowship BB/N011600/1 to I.H.M.]; Spanish Ministry of Economy and Competitiveness (MINECO) [BIO2016-79930-R to J.F.R.]; Interreg POCTEFA [EFA086/15 to J.F.R.]; European Commission [H2020 grant 676566 (MuG)].

Conflict of Interest: none declared.

References

Agius,R. *et al*. (2013) Characterizing changes in the rate of protein-protein dissociation upon interface mutation using hotspot energy and organization. *PLoS Comput. Biol*., **9**, e1003216.

Aizner,Y. *et al*. (2014) Mapping of the binding landscape for a picomolar protein-protein complex through computation and experiment. *Structure*, **22**, 636–645.

Arkadash,V. *et al*. (2017) Development of high affinity and high specificity inhibitors of matrix metalloproteinase 14 through computational design and directed evolution. *J. Biol. Chem*., **292**, 3481–3495.

Aydin,H. *et al*. (2016) Molecular architecture of the human sperm IZUMO1 and egg JUNO fertilization complex. *Nature*, **534**, 562–565.

Barlow,K.A. *et al*. (2018) Flex ddG: rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation. *J. Phys. Chem. B*, **122**, 5389–5399.

Barradas-Bautista,D. *et al*. (2017) A systematic analysis of scoring functions in rigid-body protein docking: the delicate balance between the predictive rate improvement and the risk of overtraining. *Proteins*, **85**, 1287–1297.

Berliner,N. *et al*. (2014) Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS ONE*, **9**, e107353.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Broughton,S.E. *et al*. (2016) Conformational changes in the GM-CSF receptor suggest a molecular mechanism for affinity conversion and receptor signaling. *Structure*, **24**, 1271–1281.

Brown,N.G. *et al*. (2013) Identification of the β-lactamase inhibitor protein-II (BLIP-II) interface residues essential for binding affinity and specificity for class A β-lactamases. *J. Biol. Chem*., **288**, 17156–17166.

Clark,A.J. *et al*. (2017) Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the gp120 glycoprotein of HIV-1. *J. Mol. Biol*., **429**, 930–947.

Clark,L.A. *et al*. (2006) Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci*., **15**, 949–960.

Cohen-Khait,R. and Schreiber,G. (2016) Low-stringency selection of TEM1 for BLIP shows interface plasticity and selection for faster binders. *Proc. Natl. Acad. Sci. USA*, **113**, 14982–14987.

Cunningham,B.C. and Wells,J.A. (1991) Rational design of receptor-specific variants of human growth hormone. *Proc. Natl. Acad. Sci. USA*, **88**, 3407–3411.

Dall'Acqua,W. *et al*. (1998) A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry*, **37**, 7981–7991.

Das,J. *et al*. (2014) Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. *Hum. Mutat*., **35**, 585–593.

Dehouck,Y. *et al*. (2013) BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res*., **41**, W333–W339.

Dourado,D.F. and Flores,S.C. (2014) A multiscale approach to predicting affinity changes in protein-protein interfaces. *Proteins*, **82**, 2681–2690.

Epa,V.C. *et al*. (2013) Structural model for the interaction of a designed Ankyrin Repeat Protein with the human epidermal growth factor receptor 2. *PLoS One*, **8**, e59163.

Fleishman,S.J. *et al*. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–821.

Gardsvoll,H. *et al*. (2006) Characterization of the functional epitope on the urokinase receptor. Complete alanine scanning mutagenesis supplemented by chemical cross-linking. *J. Biol. Chem*, **281**, 19260–19272.

Geng,C. *et al*. (2016) Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Eng. Des. Sel*., **29**, 291–299.

Gossage,L. *et al*. (2014) An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. *Hum. Mol. Genet*., **23**, 5976–5988.

Guntas,G. *et al*. (2010) Engineering a protein-protein interface using a computationally designed library. *Proc. Natl. Acad. Sci. USA*, **107**, 19296–19301.

Guo,W. *et al*. (2014) Hot spot-based design of small-molecule inhibitors for protein-protein interactions. *Bioorg. Med. Chem. Lett*., **24**, 2546–2554.

Horovitz,A. and Fersht,A.R. (1990) Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *J. Mol. Biol*., **214**, 613–617.

Howitt,J. *et al*. (2003) Structural basis for variation in adenovirus affinity for the cellular coxsackievirus and adenovirus receptor. *J. Biol. Chem*., **278**, 26208–26215.

Hu,J. *et al*. (2016) Conservation of hot regions in protein-protein interaction in evolution. *Methods*, **110**, 73–80.

Huang,X. and Brutlag,D.L. (2007) Dynamic use of multiple parameter sets in sequence alignment. *Nucleic Acids Res*., **35**, 678–686.

Hwang,H. *et al*. (2014) Binding interface prediction by combining protein-protein docking results. *Proteins*, **82**, 57–66.

Jemimah,S. *et al*. (2017) PROXiMATE: a database of mutant protein-protein complex thermodynamics and kinetics. *Bioinformatics*, **33**, 2787–2788.

Jemth,P. *et al*. (2014) A frustrated binding interface for intrinsically disordered proteins. *J. Biol. Chem*., **289**, 5528–5533.

Kastritis,P.L. *et al*. (2011) A structure-based benchmark for protein-protein binding affinity. *Protein Sci*., **20**, 482–491.

Kiel,C. *et al*. (2004) A detailed thermodynamic analysis of ras/effector complex interfaces. *J. Mol. Biol*., **340**, 1039–1058.

Kulkarni,M.V. *et al*. (2010) Two independent histidines, one in human prolactin and one in its receptor, are critical for pH-dependent receptor recognition and activation. *J. Biol. Chem*., **285**, 38524–38533.

Kumar,M.D. and Gromiha,M.M. (2006) PINT: protein-protein interactions thermodynamic database. *Nucleic Acids Res*., **34**, D195–D198.

Lai,J.K. *et al*. (2017) Enhancing structure prediction and design of soluble and membrane proteins with explicit solvent-protein interactions. *Structure*, **25**, 1758–1770.

Levy,E.D. (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol*., **403**, 660–670.

Li,M. *et al*. (2016) MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. *Nucleic Acids Res*., **44**, W494–W501.

Li,W. *et al*. (1997) Protein-protein interaction specificity of Im9 for the endonuclease toxin colicin E9 defined by homologue-scanning mutagenesis. *J. Biol. Chem*., **272**, 22253–22258.

Liu,Q. *et al*. (2015) Co-occurring atomic contacts for the characterization of protein binding hot spots. *PLoS One*, **10**, e0144486.

Liu,Q. *et al*. (2017). dbMPIKT: a web resource for the kinetic and thermodynamic database of mutant protein interactions. *arXiv: 1708.01857*.

Lu,S.M. *et al*. (2001) Predicting the reactivity of proteins from their sequence alone: kazal family of protein inhibitors of serine proteinases. *Proc. Natl. Acad. Sci. U.S.A*., **98**, 1410–1415.

Melo,R. *et al*. (2016) A machine learning approach for hot-spot detection at protein-protein interfaces. *Int. J. Mol. Sci*., **17**.

Moal,I.H. and Fernandez-Recio,J. (2012) SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.

Moal,I.H. and Fernandez-Recio,J. (2013) Intermolecular contact potentials for protein-protein interactions extracted from binding free energy changes upon mutation. *J. Chem. Theory Comput.*, **9**, 3715–3727.

Moal,I.H. *et al.* (2013) The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics*, **14**, 286.

Moal,I.H. *et al.* (2015a) CCharPPI web server: computational characterization of protein-protein interactions from structure. *Bioinformatics*, **31**, 123–125.

Moal,I.H. *et al.* (2015b) Inferring the microscopic surface energy of protein-protein interfaces from mutation data. *Proteins*, **83**, 640–650.

Moal,I.H. *et al.* (2017) IRaPPA: information retrieval based integration of biophysical models for protein assembly selection. *Bioinformatics*, **33**, 1806–1813.

Moretti,R. *et al.* (2013) Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins*, **81**, 1980–1987.

Nam,T.W. *et al.* (2008) Analyses of Mlc-IIBGlc interaction and a plausible molecular mechanism of Mlc inactivation by membrane sequestration. *Proc. Natl. Acad. Sci. USA*, **105**, 3751–3756.

Pallara,C. *et al.* (2013) Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges. *Proteins*, **81**, 2192–2200.

Pantazes,R.J. *et al.* (2015) The Iterative Protein Redesign and Optimization (IPRO) suite of programs. *J. Comput. Chem.*, **36**, 251–263.

Peng,Y. and Alexov,E. (2016) Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. *Proteins*, **84**, 232–239.

Petukh,M. *et al.* (2015a) On human disease-causing amino acid variants: statistical study of sequence and structural patterns. *Hum. Mutat.*, **36**, 524–534.

Petukh,M. *et al.* (2015b) Predicting binding free energy change caused by point mutations with knowledge-modified MM/PBSA method. *PLoS Comput. Biol.*, **11**, e1004276.

Pfeiffenberger,E. *et al.* (2017) A machine learning approach for ranking clusters of docked protein-protein complexes by pairwise cluster comparison. *Proteins*, **85**, 528–543.

Piepenbrink,K.H. *et al.* (2013) The basis for limited specificity and MHC restriction in a T cell receptor interface. *Nat. Commun.*, **4**, 1948.

Pires,D.E. *et al.* (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.

Plattner,N. *et al.* (2017) Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.*, **9**, 1005–1011.

Reichmann,D. *et al.* (2007) Binding hot spots in the TEM1-BLIP interface in light of its modular architecture. *J. Mol. Biol.*, **365**, 663–679.

Roisman,L.C. *et al.* (2001) Structure of the interferon-receptor complex determined by distance constraints from double-mutant cycles and flexible docking. *Proc. Natl. Acad. Sci. USA*, **98**, 13231–13236.

Rosenfeld,R. *et al.* (2017) Improved antibody-based ricin neutralization by affinity maturation is correlated with slower off-rate values. *Protein Eng. Des. Sel.*, **30**, 611–617.

Seiradake,E. *et al.* (2006) Structural and mutational analysis of human Ad37 and canine adenovirus 2 fiber heads in complex with the D1 domain of coxsackie and adenovirus receptor. *J. Biol. Chem.*, **281**, 33704–33716.

Simoes,I.C. *et al.* (2017). 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.*, **57**, 60–72.

Sirin,S. *et al.* (2016) AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Sci.*, **25**, 393–409.

Stevers,L.M. *et al.* (2017) Structural interface between LRRK2 and 14-3-3 protein. *Biochem. J.*, **474**, 1273–1287.

Sumbul,F. *et al.* (2015) Allosteric Dynamic Control of Binding. *Biophys. J.*, **109**, 1190–1201.

Thomas,C. *et al.* (2011) Structural linkage between ligand discrimination and receptor activation by type I interferons. *Cell*, **146**, 621–632.

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

Tidow,H. *et al.* (2006) Effects of oncogenic mutations and DNA response elements on the binding of p53 to p53-binding protein 2 (53BP2). *J. Biol. Chem.*, **281**, 32526–32533.

Vreven,T. *et al.* (2015) Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.*, **427**, 3031–3041.

Winn,M.D. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 235–242.

Wu,L.C. *et al.* (2002) Two-step binding mechanism for T-cell receptor recognition of peptide MHC. *Nature*, **418**, 552–556.

Xiong,P. *et al.* (2017) BindProfX: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J. Mol. Biol.*, **429**, 426–434.

Yan,Y. *et al.* (2017) Interaction entropy for computational alanine scanning. *J. Chem. Inf. Model.*, **57**, 1112–1122.

Zhao,N. *et al.* (2014) Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput. Biol.*, **10**, e1003592.