doi:10.1016/j.jmb.2011.09.031 J. Mol. Biol. (2011) 414, 289–302

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

# Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology

Sarel J. Fleishman[1], Timothy A. Whitehead[1], Eva-Maria Strauch[1], Jacob E. Corn[1], Sanbo Qin[2], Huan-Xiang Zhou[2], Julie C. Mitchell[3], Omar N. A. Demerdash[4], Mayuko Takeda-Shitaka[5], Genki Terashi[5], Iain H. Moal[6], Xiaofan Li[6], Paul A. Bates[6], Martin Zacharias[7], Hahnbeom Park[8], Jun-su Ko[8], Hasup Lee[8], Chaok Seok[8], Thomas Bourquard[9,10,11], Julie Bernauer[10], Anne Poupon[12,13,14], Jérôme Azé[10], Seren Soner[15], Şefik Kerem Ovalı[15], Pemra Ozbek[15], Nir Ben Tal[16], Türkan Haliloglu[15], Howook Hwang[17], Thom Vreven[17], Brian G. Pierce[17], Zhiping Weng[17], Laura Pérez-Cano[18], Carles Pons[18], Juan Fernández-Recio[18], Fan Jiang[19], Feng Yang[20], Xinqi Gong[20], Libin Cao[20], Xianjin Xu[20], Bin Liu[20], Panwen Wang[20], Chunhua Li[20], Cunxin Wang[20], Charles H. Robert[21], Mainak Guharoy[21], Shiyong Liu[22], Yangyu Huang[22], Lin Li[22], Dachuan Guo[22], Ying Chen[22], Yi Xiao[22], Nir London[23], Zohar Itzhaki[23], Ora Schueler-Furman[23], Yuval Inbar[24], Vladimir Potapov[24], Mati Cohen[24], Gideon Schreiber[24], Yuko Tsuchiya[25], Eiji Kanamori[26], Daron M. Standley[27], Haruki Nakamura[25], Kengo Kinoshita[28], Camden M. Driggers[29], Robert G. Hall[30], Jessica L. Morgan[29], Victor L. Hsu[29], Jian Zhan[31], Yuedong Yang[31], Yaoqi Zhou[31], Panagiotis L. Kastritis[32], Alexandre M. J. J. Bonvin[32], Weiyi Zhang[33], Carlos J. Camacho[33], Krishna P. Kilambi[34], Aroop Sircar[34], Jeffrey J. Gray[34], Masahito Ohue[35], Nobuyuki Uchikoga[35], Yuri Matsuzaki[35], Takashi Ishida[35], Yutaka Akiyama[35], Raed Khashan[36], Stephen Bush[36], Denis Fouches[36], Alexander Tropsha[36], Juan Esquivel-Rodríguez[37], Daisuke Kihara[37], P. Benjamin Stranges[38], Ron Jacak[38], Brian Kuhlman[38], Sheng-You Huang[39], Xiaoqin Zou[39], Shoshana J. Wodak[40,41,42], Joel Janin[43] and David Baker[1,44]*

[1]*Department of Biochemistry, University of Washington, Seattle, WA 98195, USA*
[2]*Department of Physics and Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306, USA*
[3]*Departments of Mathematics and Biochemistry, University of Wisconsin, Madison, WI 53706, USA*
[4]*Biophysics and Medical Sciences Training Programs, University of Wisconsin, Madison, WI 53706, USA*

[5]School of Pharmacy, Kitasato University, Tokyo, Japan
[6]Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, London, UK
[7]Physics Department, Technical University Munich, 85748 Garching, Germany
[8]Department of Chemistry, Seoul National University, Seoul 151-747, Korea
[9]INRIA AMIB, Bioinformatics Group, Laboratoire de Recherche en Informatique, Université Paris-Sud, 91405 Orsay, France
[10]INRIA AMIB, Bioinformatics Group, Laboratoire d'Informatique (LIX), École Polytechnique, 91128 Palaiseau, France
[11]INRIA Nancy/Laboratoire Lorrain de Recherche en Informatique et ses Applications, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France
[12]BIOS Group, INRA, UMR85, Unité Physiologie de la Reproduction et des Comportements, 37380 Nouzilly, France
[13]CNRS, UMR6175, 37380 Nouzilly, France
[14]Université Francois Rabelais, 37041 Tours, France
[15]Polymer Research Center and Chemical Engineering Department, Bogazici University, Bebek, Istanbul, Turkey
[16]Department of Biochemistry and Molecular Biology, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel
[17]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, USA
[18]Life Sciences Department, Barcelona Supercomputing Center (BSC), Jordi Girona 29, 08034 Barcelona, Spain
[19]Institute of Physics, Chinese Academy of Sciences, Beijing, China
[20]College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100124, China
[21]Laboratoire de Biochimie Théorique CNRS-UPR 9080, Institut de Biologie Physico-Chimique (IBPC), Paris, France
[22]Department of Physics, Huazhong University of Science and Technology, Wuhan, China
[23]Department of Microbiology and Molecular Genetics, Institute for Medical Research Israel-Canada, Hadassah Medical School, The Hebrew University, POB 12272, Jerusalem 91120, Israel
[24]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, Israel
[25]Institute for Protein Research, Osaka University, Osaka, Japan
[26]Japan Biological Informatics Consortium, Japan
[27]Systems Immunology Lab, WPI Immunology Frontier Research Center (IFReC), Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan
[28]Graduate School of Information Sciences, Tohoku University, Sendai, Japan
[29]Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR, USA
[30]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, OR, USA
[31]Indiana University School of Informatics, Indiana University Purdue University at Indianapolis, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA
[32]Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands
[33]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA
[34]Department of Chemical and Biomolecular Engineering and the Program in Molecular Biophysics, Johns Hopkins University, Baltimore, MD, USA
[35]Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan
[36]Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599-7360, USA
[37]Department of Computer Science, Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA
[38]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599-7260, USA
[39]Department of Physics, Department of Biochemistry, Dalton Cardiovascular Research Center, Informatics Institute, University of Missouri-Columbia, Columbia, MO 65211, USA
[40]Molecular Structure and Function Program, Hospital for Sick Children, Toronto, Ontario, Canada M5G 1X8
[41]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada M5S 1A8
[42]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A8
[43]IBBMC UMR 8619, Bat. 430, Université Paris-Sud, 91405 Orsay, France
[44]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

*Received 26 May 2011;*
*received in revised form*
*8 September 2011;*
*accepted 16 September 2011*
Available online
29 September 2011

***Edited by M. Sternberg***

***Keywords:***
computational protein design;
negative design;
protein–protein interactions;
conformational plasticity

The CAPRI (Critical Assessment of Predicted Interactions) and CASP (Critical Assessment of protein Structure Prediction) experiments have demonstrated the power of community-wide tests of methodology in assessing the current state of the art and spurring progress in the very challenging areas of protein docking and structure prediction. We sought to bring the power of community-wide experiments to bear on a very challenging protein design problem that provides a complementary but equally fundamental test of current understanding of protein-binding thermodynamics. We have generated a number of designed protein–protein interfaces with very favorable computed binding energies but which do not appear to be formed in experiments, suggesting that there may be important physical chemistry missing in the energy calculations. A total of 28 research groups took up the challenge of determining what is missing: we provided structures of 87 designed complexes and 120 naturally occurring complexes and asked participants to identify energetic contributions and/or structural features that distinguish between the two sets. The community found that electrostatics and solvation terms partially distinguish the designs from the natural complexes, largely due to the nonpolar character of the designed interactions. Beyond this polarity difference, the community found that the designed binding surfaces were, on average, structurally less embedded in the designed monomers, suggesting that backbone conformational rigidity at the designed surface is important for realization of the designed function. These results can be used to improve computational design strategies, but there is still much to be learned; for example, one designed complex, which does form in experiments, was classified by all metrics as a nonbinder.


## Introduction

Protein–protein interactions underlie all biological processes. Despite the availability of many co-crystal structures of complexes, our understanding of the energetics of protein association is incomplete, and this limits our ability to consistently predict the structures of complexes from monomers, predict the energetic effects of mutations at protein interfaces, and engineer high-affinity and high-specificity interactions. An improved understanding of binding energetics therefore holds the key to resolving some of the most important problems in protein biophysics and molecular biology.

A recently developed method for de novo binder design produced two proteins that interacted with a sterically hindered surface on Spanish influenza hemagglutinin (SC1918/H1 HA; hereafter referred to as HA).[1] Following *in vitro* evolution, two to four mutations in the periphery of each of these interfaces improved binding to low nanomolar dissociation constants and one of the proteins inhibited HA function. However, 71 other designed proteins which expressed robustly in yeast cell surface display experiments[2] and were predicted to bind did not experimentally interact with HA. The Baker group has had similar low success rates with other de novo interface design problems (to be published), highlighting limitations in the understanding of protein-binding energetics and their repercussions for the ability to design novel protein functions. More sensitive experimental detection methods could identify additional binders in this set (the current method requires dissociation constants better than 10 μM and binding off-rates less than 10 $s^{-1}$), but the ability to computationally generate high-affinity interactions is vital for engineering new protein functions.

We asked the protein-docking community to help identify what was missing in our protein-modeling calculations. This article describes the benchmark tests we established and summarizes the insights from the many interface-modeling experts who took up the challenge.

*Corresponding author. Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.
E-mail address: dabaker@uw.edu.

Present address: S. J. Fleishman, Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel; J. E. Corn, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA.

Abbreviations used: CAPRI, Critical Assessment of Predicted Interactions; PDB, Protein Data Bank; ROC, receiver operator characteristic; AUC, area under the curve.

## Results

### A protein-interface design benchmark

The computational interface design protocol consists of (i) pre-computing a set of high-affinity amino acid residue interactions with the target surface, (ii) redesigning natural protein scaffolds to incorporate a number of these amino acids, and (iii) designing the remainder of the interface to enhance binding affinity.[1] This protocol can produce protein complexes with computed binding characteristics that rival natural complexes. For instance, the distributions of interface buried surface areas and computed binding energies of designed and naturally occurring protein complexes overlap (Fig. 1; Table S1). In many cases, designed protein complexes show more favorable values than do natural complexes. This is despite the fact that the vast majority of the designed complexes do not experimentally bind. The discrepancy between prediction and experiment is the focus of this study: our goal is to identify the missing components in binding-energy calculations to improve both our ability to design high-affinity interfaces and, more generally, our understanding of protein-association thermodynamics.

We set out to identify thermodynamic components of binding that are poorly modeled and could be the underlying cause of the low success rate of de novo binder design. In a preliminary experiment, a set of 20 designed binders of several targets that did not show detectable binding to their targets was provided to participants in the community-wide experiment on the Critical Assessment of Predicted Interactions (CAPRI),[3] alongside one experimentally determined but, at that time, unpublished co-crystal structure of two proteins that bound with a low-nanomolar dissociation constant.[4] The participants were asked to rank the 21 complexes according to their propensity to bind in the modeled or experimentally determined binding mode. In this preliminary experiment, only 2 of 28 participating groups (Groups 1 and 6, Table 1) clearly identified the co-crystal structure as the true binder—a performance that is not significantly different from chance at 5% confidence (to be discussed in the next Special Issue on CAPRI). The successful groups relied on metrics that were largely based on electrostatics calculations. Notably, the Rosetta energy function, which was used in the design process, explicitly treats hydrogen bonding and solvation, but because of difficulties in accurately modeling long-range electrostatics interactions does not attempt to model these explicitly.[5] These results suggested that the task of identifying complexes that are likely to bind is nontrivial and that a larger-scale community-wide investigation could provide considerable insight into this problem.

To set up a benchmark for a more comprehensive community-wide investigation into the elements that are missing in our evaluation of binding thermodynamics, we prepared a set of 87 designed proteins targeting three different proteins of interest (models are available as Supplemental Data, and plasmids encoding genes for expressing the designs using yeast cell-surface display are available†). The three target proteins were Spanish influenza HA [62% of the designed complexes; chains A and B of Protein Data Bank (PDB) entry 3GBN[6]], the acyl-carrier protein 2 from *Mycobacterium tuberculosis* (25%; *Mt* ACP2; PDB entry 2CGQ), and the Fc region of human IgG1 antibodies (13%; PDB entry 1L6X[7]). The structures of the scaffold proteins for binder design were taken from the PDB, and their surfaces were redesigned for binding using the computational method mentioned above.[1] As a reference set of solved co-crystal structures, we used the docking benchmark 3.0[8] comprising 120 protein complexes with experimentally determined dissociation constants[9] ranging from $10^{-5}$ to $10^{-14}$ M. These sets of natural and computationally designed complexes were provided to participants in CAPRI, noting in each case whether a complex was designed or natural. At the beginning of the experiment, nine designed proteins had not been experimentally tested for binding and these served as unmarked blind cases.

Each participating group (Table 1) was asked to provide a method for ranking the complexes according to their binding energy (all of the values provided by participants are available as Supplemental Data). To get at the underlying physical chemistry of binding, we asked groups not to train their methods on the data; that is, the information on whether a complex was designed or natural could not be used in training the parameters used in the evaluation strategy. Otherwise, the groups were free to choose which metrics or combinations of metrics to use. Figure 2 shows a receiver operator characteristic (ROC) curve for each participating group, plotting the true-positive rate *versus* the false-positive rate. The area under the curve (AUC, in percentage) is marked in each panel. The participating groups were additionally asked to categorize each complex according to the following criteria: the two partners (i) bind, (ii) are likely to bind, (iii) are likely not to bind, (iv) do not bind, and (v) unknown (Fig. S1). They were also free to choose thresholds to maximize discrimination.

The methods used by participating groups span a wide spectrum. Many groups computed binding energies, typically dominated by electrostatics, solvation, and knowledge-based pair terms (Groups 1, 5, 6, 11, 12, 14, 20, 23, 26, 28, 29, 31, 33, and 36); Groups 1 and 6 used continuum solvation methods to compute binding energies, similar to widely used molecular mechanics/Poisson–Boltzmann surface

† http://www.addgene.com

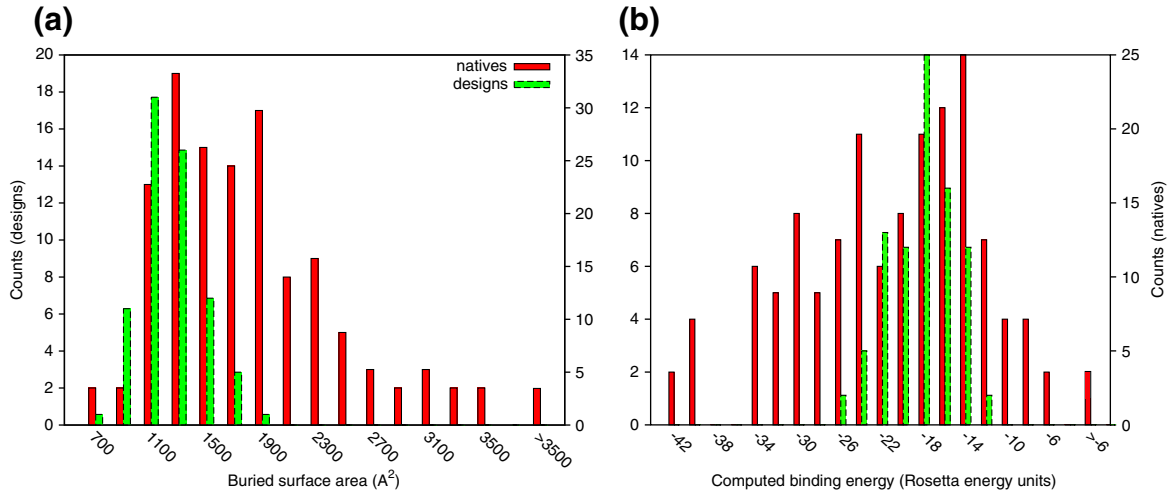


**Fig. 1.** Natural and designed complexes have similar overall properties. (a) Buried surface area at the interface; (b) computed binding energy. Computed binding energies are reasonably correlated with experimentally determined dissociation constants (Pearson correlation $r = 0.53$; Ref. 15). All plots were produced using gnuplot 4.4 and enhanced with Adobe Illustrator. In all figures, native refers to natural complexes in the docking benchmark.[8]

area approaches for computing binding affinities.[10] Others utilized features such as hydrogen-bonding patterns and buried surface area (Groups 16, 21, 23, 24, 30, 32, and 35). Groups 2 and 22 used machine learning to determine which features discriminate previously published Rosetta models from natural complexes. Groups 8 and 17 used the sequence conservation at the protein interface as a discriminator. Group 10 analyzed low-frequency dynamics, and Group 7 tested the low-resolution compatibility of the surfaces compared to randomly docked decoys of the same partners.

## Discrimination between the designed interfaces and some, but not all, categories of natural ones

Many different metrics provide useful *posteriori* discriminators between designed and naturally occurring complexes (Fig. S1), with several groups achieving AUC values above 85% (Fig. 2). However, the ROC curves also point out that even well-performing metrics suffer from poor discrimination between designs and many native complexes; many of the best discriminators rank a large fraction of the natural complexes as better binders than the designed complexes but still rank many designed and natural ones equally. Consequently, many of the native complexes were predicted as unlikely to bind or as not binding by most groups. These results suggest that the designs share some features with a substantial fraction of the natural complexes but not with all.

To get a more detailed view of the individual features that contribute most to discrimination, we compared the distributions for designed and natural interfaces of the two most heavily weighted terms given by several participating groups (Fig. 3a). As with the full metrics (Fig. 2 and Fig. S1), the individual-score values for natural complexes span and exceed the range of designed complexes, and hence, no single or indeed pair of scores unambiguously discriminates designed from natural complexes. Nevertheless, the designed complexes typically stand out as having, on average, less optimal values than a majority of the natural complexes in terms of their van der Waals contacts, solvation self-energy, and electrostatic complementarity. To understand the commonalities between designed and natural complexes that were predicted not to bind, we analyzed in detail the results from Group 6, one of the best-performing participants (Fig. 2). We found that those natural interfaces that scored more favorably than designs according to the two-metric analysis (Fig. 3a) were typically larger and contained many salt bridge or backbone-mediated interactions (see per-group two-metric analysis in Supplemental Data). By contrast, the natural interfaces that were predicted not to bind were smaller, more hydrophobic, and contained few, if any, charges and paired backbone atoms. The de novo designed interfaces share many of the same features as the latter category of smaller, more hydrophobic natural interfaces, explaining why many metrics showed natural complexes to span the range of values for the designs but did not clearly discriminate the two groups (Figs. 2 and 3a). Many of these natural hydrophobic protein complexes bind quite strongly, implying that even the best-performing metrics do not fully reflect binding thermodynamics. This is highlighted by the fact that the natural complex best separated from the designs (predicted most strongly to be a binder) was a structure, which after its publication was deemed by several studies to be likely incorrect,[11] and was

**Table 1.** List of participating groups and a brief explanation of the methods

| Group number[a] | Affiliation[b] | van der Waals packing[c] | Solvation[c] | Pair terms[c] | Electrostatics[c] | Others[c] | Use of prior knowledge[d] |
|---|---|---|---|---|---|---|---|
| 1 | 2 | | | | 1 | Electrostatic interaction free energy, calculated on the transient complex, by solving the Poisson–Boltzmann equation | a |
| 2 | 3, 4 | NA | NA | NA | NA | Support-vector machine (SVM) | b |
| 5 | 5 | | | 1 | | | a |
| 6 | 6 | 0.1 | 0.4 | 0.16 | | | c |
| 7 | 7 | — | — | ATTRACT score of the minimized complex (0.33) in RT units | — | Rank of minimized complex relative to docking solutions from systematic search (0.33); deviation of complex from nearest minimum (0.33) | a |
| 8 | 8 | | | 0.18 | | Sequence conservation score (0.52) | a |
| 9 | 9–14 | NA | NA | NA | NA | Side-chain entropy (0.13) Genetic algorithms | a |
| 10 | 15, 16 | | | | | DiffColl (1.0) | a |
| | | | | | | The difference in the increase in degree collectivity between chains A and B | |
| 11 | 17 | 0.41 (van der Waals attractive) | 0.42 | | 0.13–0.21 (four independent weights for short/long/attractive/repulsive; average is 0.16) | | c |
| 12 | 18 | | 0.5 | | 0.5 | | a |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 14 | 19 | 0.056 | Sum of weights for 18 terms of DeLisi–Zhang atomic solvation (0.563) | Dfire (0.369) | 0.013 | Sum of linear fitting weights for DeLisi–Zhang atomic solvation (4.101), for pair solvation and hydrogen bond (−1.167), and for many-body graph (−3.712) | | a |
| 16 | 20 | | 0 | | 0 | rpscore3 (1.0) | 0 Interface area ≥1200 Å$^2$; Interface patch analysis1 | a |
| 17 | 21 | — | — | — | — | Relative sequence entropy score comparing the degree of conservation of the interface core *versus* the rim (1.0) | | a |
| 20 | 22 | 0 | 0 | 1 | 0 | 0 | | a |
| 21 | 23 | NA | NA | NA | NA | Interface descriptors: polar solvent-accessible surface area buried at the interface is smaller in designs | | c |
| 22 | 24 | 0.2 | 0 | 0.2 | 0.2 | 0.2 | | b |
| 23 | 25–28 | 0.09 | 0.28 | | 0.44 | SCRsurf<br>−0.19 | | a<br>a |
| 24 | 29, 30 | NA | NA | NA | NA | Interface intra- and intermolecular energies scaled to differentiated total energy, and scaled surface area buried at the interface | | a |
| 26 | 31 | 0 | 0 | 1 | 0 | 0 | | a |
| 28 | 32 | 0.3 | 0.26 (Lazaridis–Karplus solvation+ buried surface area) | | 0.24 | Hydrogen bonding<br>0.2 | | a |

(*continued on next page*)

**Table 1** (*continued*)

| Group number[a] | Affiliation[b] | van der Waals packing[c] | Solvation[c] | Pair terms[c] | Electrostatics[c] | Others[c] | Use of prior knowledge[d] |
|---|---|---|---|---|---|---|---|
| 29 | 33 | 0.25 | 0.25 | | 0.25 | Internal energy | a |
| 30 | 34 | NA | NA | NA | NA | Interface area per residue of the complex | c |
| 31 | 35 | 0.75 | 0.05 | | 0.2 | | a |
| 32 | 36 | NA | NA | NA | NA | Frequency and geometric similarity of interaction patterns of interfacial residues to the native (classical) ones | a |
| 33 | 37 | van der Waals attractive (0.49) | 0.01 | 0.35 | Short range attractive (0.06) | Hydrogen bonding | a |
| 35 | 38 | NA | NA | NA | Short range repulsive (0.07) NA | Binding energy (dG) per surface area (PSA) (0.25), hydrogen-bond energy per dG (0.25), cavity score (0.25), unsatisfied hydrogen bonds PSA (0.25) | a |
| 36 | 39 | NA | NA | ITScore/PP | NA | NA | a |

A complete description of each method is provided in Supplemental Methods.

[a] The group number refers to the numbers in the main text and figures.

[b] The affiliation number in the author affiliation section.

[c] Weights on the major score terms used by the discriminators all terms including minor ones are listed in Supplemental Methods. The weights in the table are reported after normalizing the sum of all weights used by each group to 1.0.

[d] Extent to which prior knowledge was used: (a) none; (b) score was trained on Rosetta models provided in the past, but not on the design benchmark; (c) different discrimination models or parameters were tested and the best-performing one was selected.

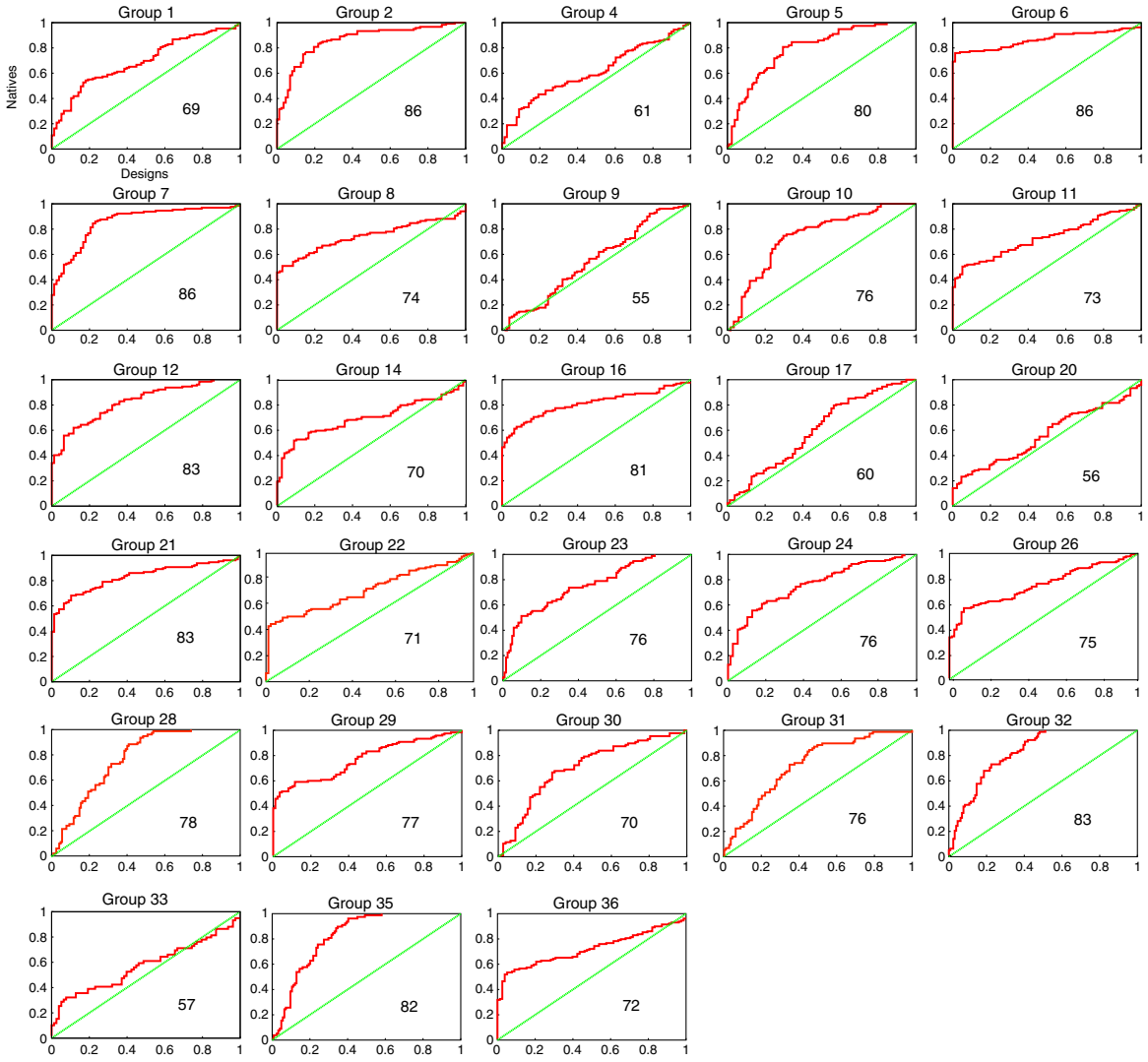


**Fig. 2.** Ability of different methods to discriminate between native and designed complexes. ROC curves are shown for each group, with the true- and false-positive classification on the *y*- and *x*-axes, respectively. The steeper the ascent of the curve and the larger the AUC, the better the discrimination between natural and designed complexes. The green diagonal represents the expected output of random prediction. Percent AUC is noted within each plot. Groups 2 and 22 trained their metrics, in part, on Rosetta models published in the past but not on the current set of designs (see Supplemental for more details).

recommended for retraction by the University of Alabama (PDB entry: 1BGX[12]). In retrospect, the bias towards hydrophobic interfaces was a failing of our design benchmark set. We remedied this failing in two ways (below): by adding more polar interfaces to the design set and by contrasting the designs with the most apolar natural interfaces in the docking data set.

### Reducing the polarity discrepancy between natural and designed interfaces identifies methods that discriminate designs based on binding site rigidity

To address the problem of unequal polarities in designed and natural interfaces, we reoptimized the sequences of 87 designed complexes, increasing the contributions from residue pairwise-interaction probabilities and Coulomb electrostatics to the energy function used by RosettaDesign, and selected 29 designs with high buried surface area and favorable binding energies. In these redesigned interfaces, the distributions of contributions to binding from electrostatic and pairwise-interaction probabilities are comparable to those of natural interfaces (Fig. 3b). While these new redesigned complexes have many flaws (side-chain packing is not ideal and their interfaces contain many unsatisfied hydrogen-bond donors and acceptors), the addition of interfaces with higher charge complementarity reduces the polarity discrepancy between

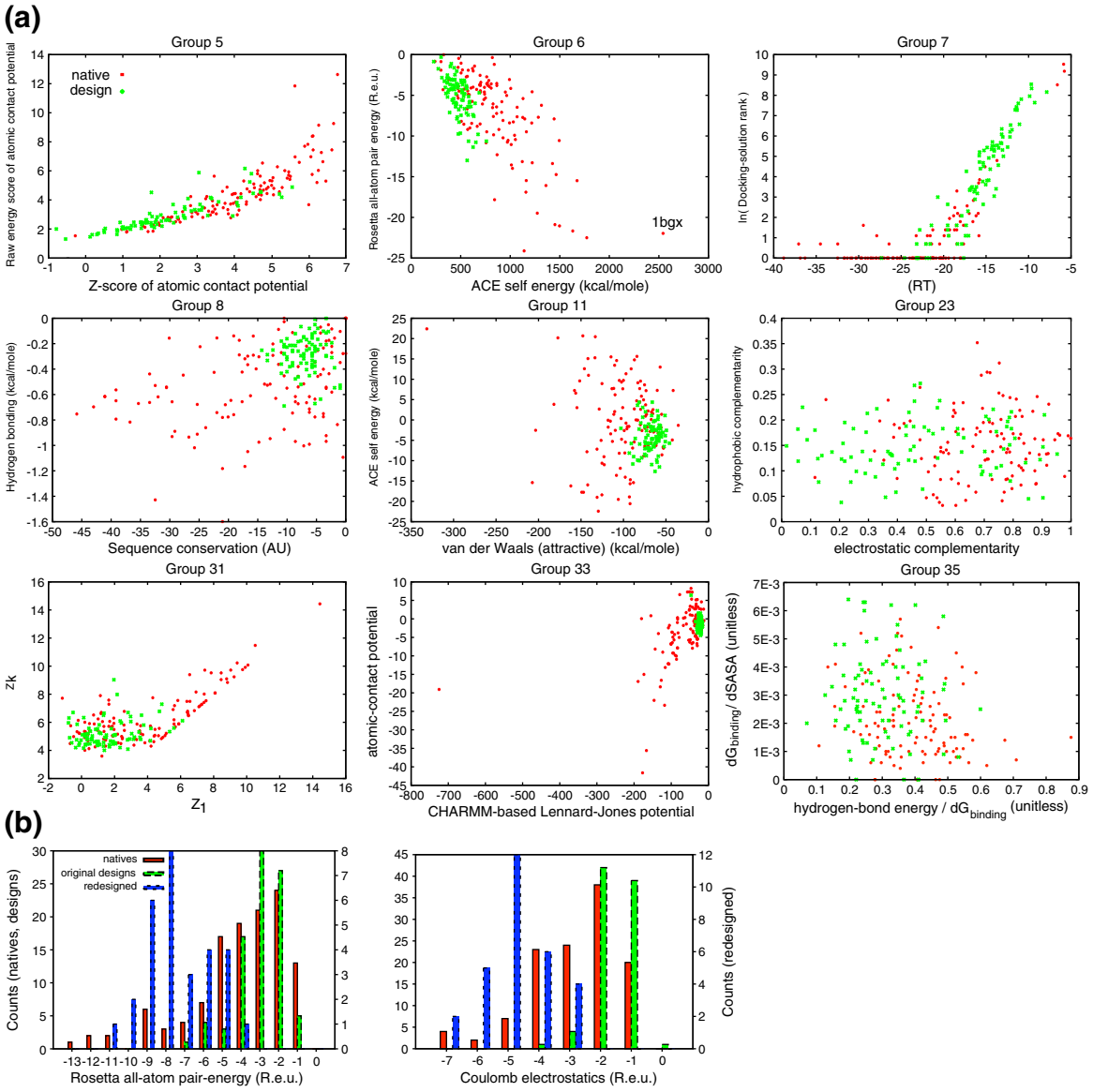


**Fig. 3.** Individual features that partially discriminate native and designed complexes. (a) Comparison of natives and designs using the two most heavily weighted terms in the scoring function for each group. The points represent individual natives or designs, and the axes represent the most heavily weighted scoring terms. The scatter plots provide insight into some of the discriminatory power of the methods. While the phase planes occupied by designs and natives overlap, in these cases, the designs occupy a small fraction of the plane with many of the natives having more favorable values. The results from Groups 11 and 33 suggest that the van der Waals contacts in designed interfaces are weaker than those in natives. Likewise, Groups 6 and 11 suggest that solvation self-energy (ACE) and electrostatics (the dominant contribution to Rosetta pair energy) are more optimized in natives. See individual groups' methods for more details. (b) Modification of the design protocol yields distributions of interface pairwise and Coulomb electrostatic energies similar to those in natural complexes. Natural complexes (natives) and designs generated with (redesigns) and without (designs) an increased pairwise attractive term (weight = 0.98) and Coulomb electrostatic interaction with a distance-dependent dielectric (weight = 1.0). The distributions were calculated using a pairwise attractive term and an electrostatic interaction of 0.49 and 0.25, respectively, for all complexes. These designs have many flaws as potential binders but can serve as decoys with more native-like distributions of electrostatic interactions.

designed and natural interfaces in our set and makes the benchmark more representative of the physical–chemical diversity of natural interfaces. We have added these new, more polar complexes to the benchmark set (Supplemental Data). The improved benchmark set should provide an even better test of current understanding of binding physical chemistry than the original set.

To isolate metrics that discriminate the designs from a set of apolar natural interfaces, we selected the 25 natural interfaces with the lowest electrostatic desolvation penalty according to the Rosetta all-atom energy (Table S2). As expected, the AUC of many of the metrics deteriorated in this analysis compared to the results of Fig. 2, while a few methods performed as well on this stricter test as in the one shown in Fig. 2 (Table S3). Group 7 (AUC = 81% in this analysis) used low-resolution docking and favored those complexes where close-to-native conformations had lower interaction energies than far-from-native ones. An analysis of the worst- and best-performing designs according to this method showed that it penalized designs with poor low-resolution shape complementarity and conversely favored designs with intricate "knobs-into-holes" features, which allow more residue-to-residue interactions. Group 10 (AUC = 79% in this analysis) used a single feature based on the compatibility of the low-frequency vibrational modes of the partner proteins. Interfaces where the vibrational modes of the two partners were incompatible were penalized. An analysis of the worst-performing designs according to this method showed that it penalized designs where the binding surface was positioned on loops or secondary structural elements that were poorly embedded in the designed monomer and conversely favored interfaces that integrated the designed surface through many interactions in the host monomer. Group 10 found that a simpler related metric based on the average degree of connectivity of interfacial residues on the designed monomer (see Computational Methods) performed more poorly than the analysis of vibrational modes but was also discriminatory. Indeed, in following up on the Group 10 results, we found that most designed proteins with an average degree of less than 8.5 residue neighbors at the interface (~15% of designs in the set) utilize loops or secondary structural elements that are poorly anchored to the designed protein and, retrospectively, are unlikely to form the modeled surfaces in experiment (Fig. 4). That such a high

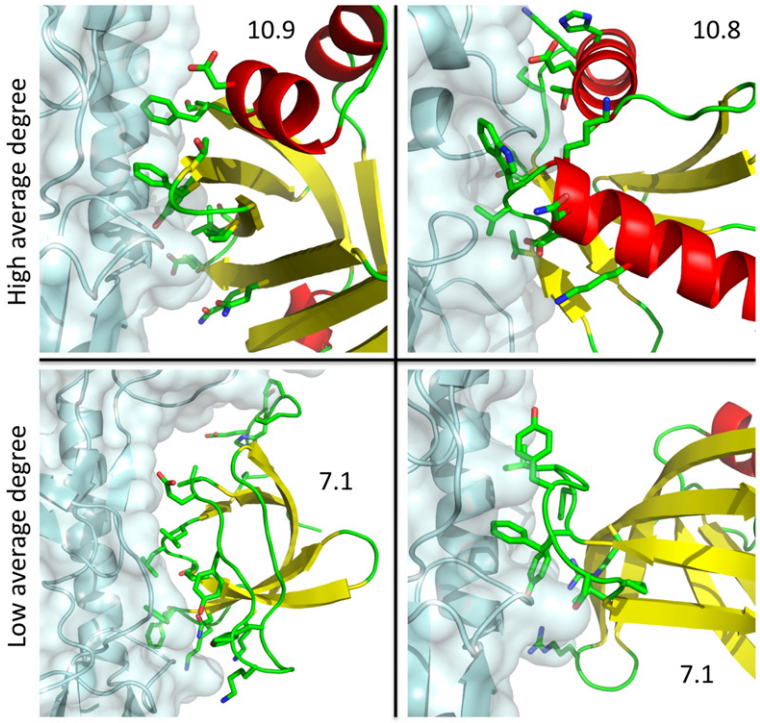


**Fig. 4.** Average number of neighbors (average degree) of interface residues within the designed monomer discriminates some designed complexes from native complexes. Surfaces with low average degree (bottom) tend to comprise segments, including unstructured regions, which are poorly embedded in the host monomer. By contrast, surfaces with high average degree (top) comprise secondary structural elements and short loops that are better structurally connected to the host monomer. Following sequence design, poorly connected surfaces might have altered conformations from those seen in the wild-type protein structure, providing some explanation for the failure of these designs to experimentally bind their targets. Average degree is marked on each panel. Clockwise from top left, the panels represent designs 47, 59, 78, and 77 (coordinates are available in the online supplement). The target proteins are rendered in cyan. The backbones of the designed monomers are colored according to secondary structure (red, helix; yellow, strand; green, loop). Designed interfacial residues are shown in sticks with carbon, oxygen, and nitrogen, colored green, red, and blue, respectively. Molecular representations were produced with PyMOL.[20]

fraction of designs employ backbones that are poorly anchored in the designed monomer is unsurprising given that binding to a target surface is typically hindered by other surfaces on the target molecule; designed surfaces that are less embedded in their host monomers suffer less from such hindrance. We have implemented this degree of connectivity metric in the Rosetta software and expect it to improve the likelihood of obtaining active designed binders in the future.

### Failure to identify an experimentally validated designed binder as such

Of the 87 designed interfaces provided to participants for ranking, 9 designs had not been tested for binding at the start of the experiment and thus serve as a blind test of the ranking methods. Of these nine, one has been experimentally confirmed to bind its HA target surface (herein numbered design 45 or HB80 in Ref. 1). *In vitro* selection of design 45 variants for higher affinity identified four substitutions at the periphery of the interface that together produced an experimentally determined dissociation constant of 38 nM, rivaling many of the affinities in the docking benchmark of naturally occurring binders.[8] Despite this high affinity, none of the groups predicted that design 45 binds, and a majority predicted that it is unlikely to bind or that it would not bind (Fig. S2). Design 45 has a small nonpolar interface, which, as noted above, confounds discrimination of binders from nonbinders by most of the methods reported here. The failure with design 45 and the general difficulty in distinguishing the designs from nonpolar natural interfaces suggest that considerable work remains in refining models of protein-interface thermodynamics.

## Discussion

Defining the structural and energetic determinants of high-affinity binding is crucial for our understanding of protein-interaction networks and the ability to intervene in physiologically important systems. Our analysis provides a snapshot of current understanding of binding energetics. While certain features emerge as discriminators between designs and a majority of the natural protein complexes in our data set, all of the metrics misclassify some natural complexes as nonbinders. In many areas of computational biology, ranging from sequence alignment[13] to function annotation,[14] the availability of comprehensive benchmarks has provided strong impetus to method development and a powerful means of gauging progress. The benchmark provided here, the first to contain complexes that are predicted to associate but have been experimentally determined not to interact, provides a valuable orthogonal axis for evaluating both the relative and absolute performance of alternative approaches.

The design discrimination test is complementary to traditional docking tests. In this test, large-scale sampling of rigid-body or backbone freedom is not needed, allowing more direct focus on the energy function. On the other hand, it must be kept in mind that the failure of a computational design to experimentally bind its target could be related not only to overestimation of the computed binding energy due to energy function inaccuracies but also to imperfect design at the monomeric protein level: the design may not actually fold to the target structure. The high likelihood of designed side chains to adopt binding-incompatible conformations in the unbound state has been suggested to play a role in the failure of design calculations to produce active binders.[15] Here, we find that changes to backbone structure in designed surfaces might play an equally significant role in compromising designs. Indeed, in the design of hemagglutinin binders, the two active designs used largely helical and conformationally restricted surfaces.[1] Our conclusion that surfaces that are not well anchored are poor choices for design can be easily used to eliminate such surfaces from design.

The 28 participating groups found many differences between the designed and natural complexes. In particular, several metrics employing electrostatics and solvation show promise as discriminators, perhaps unsurprisingly, given that the three surfaces targeted in the design set were largely hydrophobic, whereas natural interfaces span the range of hydrophobicity and charge. On the other hand, most all-atom metrics fail to discriminate natural and designed hydrophobic interfaces, even though most of the designs do not bind. This result underscores the importance of developing improved force fields for protein interfaces that are able to discriminate binders from nonbinders in all categories. One result of the community-wide testing is that our original benchmark set could be "tricked" because of its very strong focus on nonpolar interfaces. We have now supplemented the benchmark with more polar and charged interfaces to remedy this deficiency and by suggesting a subset of 25 apolar natural interfaces for comparison to designs; we look forward to the improved metrics that will be developed to solve the discrimination problem posed by this more inclusive benchmark.

Solving the discrimination problem by all-atom methods may require explicit treatment of the various conformational-entropy penalties of binding, such as side-chain and backbone freezing.[15,16] Additional aspects such as water molecules at the interface and the likelihood that the designed protein adopts its target conformation may also need to be addressed. The availability of a

comprehensive data set should enable the development of improved energy functions, yielding a more complete understanding and formulation of the energetic contributions to binding free energy and increasing the reliability of tools for predicting and engineering protein interactions.

## Materials and Methods

Experimental materials and methods and the computational methods used in discrimination are provided in the online supplement.

### Computational methods

#### *Preparation of input files*

Designed and natural complexes were subjected to the same computational protocol consisting of full side-chain repacking and refinement of the rigid-body and side-chain conformations using the local-refine mode of RosettaDock.[17] All calculations were conducted in the Rosetta all-atom force field (score12), which is dominated by van der Waals, hydrogen bonding, and solvation terms.[5] A RosettaScript for complex-structure refinement is available in the online supplement. Refined structures were provided to the participants and are available in the online supplement.

#### *Computed binding energy and buried surface area calculations*

The binding energy and buried surface area (Fig. 1; Table S1) were computed within the Rosetta software suite. For the natural complexes, the biologically relevant interface was extracted from information provided with the docking benchmark.[18] Binding-energy calculations (using score12) were computed by subtracting the energy in the unbound complex from the energy in the bound complex, in each state allowing for repacking of interface side chains. Binding energies were averaged over three repeats for numerical stability. A RosettaScript for computing the binding energies and buried surface areas is available in the online supplement.

#### *ROC and the AUC*

The raw scores from each group were numerically sorted from high to low propensity to bind, irrespective of the type of complex (natural or designed). For each natural complex in the sorted list, a step was taken along the $y$-axis to plot the ROC, and conversely, for each designed complex, a step was taken along the $x$-axis. Step sizes were normalized such that the total lengths of the $x$- and $y$-axes were 1.0. The AUC was computed by summing the area added under the curve for each $x$-axis increment. Scripts for computing the AUC and plotting the ROC are available in the online supplement.

#### *Degree of connectivity at the interface*

For each interface residue on designed monomers and all interface residues on natural binders, we calculate the number of residue neighbors on the host monomer within 8 Å of the interfacial residue (ignoring the partner protein). We find that below 8.5 residue neighbors, designed surfaces are poorly anchored in their host monomers (examples in Fig. 4). Residues within 8 Å of the partner protein were considered to be interfacial. This metric is implemented in RosettaScripts[19] (see Supplemental Data).

#### *Redesign for improved electrostatics*

The 87 designed complexes served as starting structures for three iterations of side-chain design of scaffold interface residues followed by minimization of rigid-body, backbone, and side-chain degrees of freedom. During design and minimization, the Rosetta all-atom force field was augmented with a Coulomb electrostatic interaction term with a distance-dependent dielectric (weight = 1.0) and pair potential (weight = 0.98, compared to 0.49 in the default all-atom force field). The 29 designs burying the highest surface areas were selected.

Pairwise and electrostatic contributions to binding (Fig. 3b) were these energetic components of binding-energy calculations (see above) and were computed assuming weights of 0.49 for the pairwise potential and 0.25 for Coulomb electrostatics. A RosettaScript for the design trajectory is available as Supplemental Data.

#### *Source code*

The Rosetta software suite is available online free of charge to academic users‡. Scripts used in analyzing the data and producing the graphics are provided in the online supplement.

Supplementary materials related to this article can be found online at doi:10.1016/j.jmb.2011.09.031

## Acknowledgements

The authors thank Sameer Velankar and Marc Lensink for their help in coordinating this experiment and Raik Grunberg for many helpful suggestions on a draft. S.J.F. was supported by a long-term fellowship from the Human Frontier Science Program. S.J.W. is Canada Research Chair Tier 1, funded by the Canadian Institutes of Health Research. Research in the Baker laboratory was supported by the Howard Hughes Medical Institute, the Defense Advanced Research Projects Agency, the National Institutes of Health Yeast Resource Center, and the Defense Threat Reduction Agency.

## References

1. Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M. *et al.* (2011). Computational design of proteins targeting the

‡ http://www.rosettacommons.org

conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–821.
2. Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M. & Wittrup, K. D. (2006). Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1**, 755–768.
3. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S. *et al.* (2003). CAPRI: a critical assessment of predicted interactions. *Proteins*, **52**, 2–9.
4. Karanicolas, J., Corn, J. E., Chen, I., Joachimiak, L. A., Dym, O., Peck, S. H. *et al.* (2011). A de novo protein binding pair by computational design and directed evolution. *Mol. Cell*, **42**, 250–260.
5. Das, R. & Baker, D. (2008). Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382.
6. Ekiert, D. C., Bhabha, G., Elsliger, M. A., Friesen, R. H., Jongeneelen, M., Throsby, M. *et al.* (2009). Antibody recognition of a highly conserved influenza virus epitope. *Science*, **324**, 246–251.
7. Idusogie, E. E., Presta, L. G., Gazzano-Santoro, H., Totpal, K., Wong, P. Y., Ultsch, M. *et al.* (2000). Mapping of the C1q binding site on rituxan, a chimeric antibody with a human IgG1 Fc. *J. Immunol.* **164**, 4178–4184.
8. Hwang, H., Pierce, B., Mintseris, J., Janin, J. & Weng, Z. (2008). Protein–protein docking benchmark version 3.0. *Proteins*, **73**, 705–709.
9. Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M. & Janin, J. (2011). A structure-based benchmark for protein–protein binding affinity. *Protein Sci.* **20**, 482–491.
10. Gilson, M. K. & Zhou, H. X. (2007). Calculation of protein–ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21–42.
11. Sheffler, W. & Baker, D. (2009). RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci.* **18**, 229–239.
12. Murali, R., Sharkey, D. J., Daiss, J. L. & Murthy, H. M. (1998). Crystal structure of Taq DNA polymerase in complex with an inhibitory Fab: the Fab is directed against an intermediate in the helix-coil dynamics of the enzyme. *Proc. Natl Acad. Sci. USA*, **95**, 12562–12567.
13. Thompson, J. D., Plewniak, F. & Poch, O. (1999). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
14. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
15. Fleishman, S. J., Khare, S. D., Koga, N. & Baker, D. (2011). Restricted sidechain plasticity in the structures of native proteins and complexes. *Protein Sci.* **20**, 753–757.
16. Grunberg, R., Nilges, M. & Leckner, J. (2006). Flexibility and conformational entropy in protein–protein binding. *Structure*, **14**, 683–693.
17. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A. & Baker, D. (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331**, 281–299.
18. Chen, R., Li, L. & Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, **52**, 80–87.
19. Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E. M., Khare, S. D. & Koga, N. (2011). RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One*, **6**, e20161.
20. DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* DeLano Scientific, Palo Alto, CA.