Supplemental Materials and Methods

Experimental assays

Availability of plasmids

All of the designs reported here, except for 11 that showed binding activity (designs 3, 6, 14, 22, 23, 42, 54, 57, 67, 79, and 84), are available as DNA plasmids from the AddGene service (http://www.addgene.org). Designed genes were subcloned between Nde/XhoI sites in an in-house yeast display plasmid[1], named pETCON. pETCON is the pCTCON plasmid reported in ref. [1] with the following modifications: (a) a frameshift mutation in the CD20 encoding region; (b) a Nde restriction site immediately downstream of the NheI site; and (c) a XhoI-Gly_2 spacer sequence immediately upstream of the BamHI restriction site.

Target protein preparation

Production and purification

SC1918/H1 Hemagglutinin was produced and biotinylated according to previous reports[2]. The gene encoding *Mtb* ACP2 was custom ordered from Genscript (Piscataway, NJ) and subcloned into a pET vector with C-terminal hexa-histidine and AviTags. The plasmid was transformed into BL21 (DE3) *E. coli* and protein expression was induced by the Studier autoinduction method[3].

After expression for 16 h at 18°C, cells were pelleted, resuspended into buffer HKGlu (20 mM Hepes, 150 mM potassium glutamate pH 7.4) and sonicated to lyse cells. After clarification by centrifugation, cells were applied to a gravity flow Ni^{2+}-NTA column and purified by step elution in a buffer containing 400 mM imidazole, 20 mM Hepes, 150 mM potassium glutamate pH 7.4. *Mtb* ACP2 was then desalted into buffer HKGlu, quantified by A_{280} absorbance using the estimated extinction coefficient e=12,660 $M^{-1}cm^{-1}$, and flash frozen in liquid nitrogen until further use. Mass spectrometry showed that the protein was produced full-length in the apo-form without the phosphopantetheine prosthetic group attached to Ser38. Circular dichroism scans at 22°C in PBS pH 7.4 showed pronounced minima at 208 and 222 nm indicative of a helical protein, as expected.

To test binding to the constant region of human IgG Fc, the commercially available therapeutic antibody Rituximab was used.

Biotinylation

AviTag *Mtb* ACP2 was C-terminally biotinylated using the BirA kit from Avidity (Aurora, CO). Aliquots were thawed on ice and added to a final concentration of 40 μM to a mixture containing (per 100 μL) 10 μL biomix A, 10 μL biomix B, 10 μL d-biotin (500 μM stock), 1 μL *E. coli* biotin ligase (3 mg/mL stock), and the balance buffer HKGlu. After incubation at 22°C for 5 h, *Mtb* ACP2 was separated from the biotin ligase by Ni^{2+}-NTA affinity chromatography and desalted into buffer HBS (20 mM Hepes, 150 mM sodium chloride pH 7.4) using a large-bore desalting column. Biotinylation was confirmed by mass spectrometry, and protein was flash frozen until further use.

Amino groups of Rituximab were nonspecifically biotinylated using the Chromalink Biotin labeling kit (Solulink, San Diego) following the manufacturer's instructions. Antibodies were cross-linked to 3-4 biotin molecules as indicated by absorbance measurements. Protein was stored at 4°C and used within 2 months after biotinylation.

Conjugation

Biotinylated *Mtb* ACP2 aliquots were thawed on ice and added in excess at a 6:1 molar ratio to streptavidin-phycoerythrin (SAPE) (Invitrogen, Carlsbad, CA) on ice for 1 hr. ACP2-SAPE conjugate was separated from free ACP2 on a Sephacryl S-200 size exclusion column using a flow rate of 1 mL/min and a mobile phase of buffer HBS. Conjugate was concentrated using a 100 kDa MWCO Amicon Ultra centrifugal filter unit (St. Louis, Mo.) and quantified by fluorescence intensity of PE (excitation wavelength 495 nm, emission wavelength 575 nm, cutoff 530 nm) using a Spectramax $M5^e$ fluorescence plate reader (Molecular Devices, Sunnyvale, Ca.). Unconjugated SAPE was used for the standard curve for protein quantification. Formation of ACP2-SAPE conjugate was verified by denaturing gel electrophoresis. ACP2-SAPE conjugate was diluted to 2 μM in buffer HBS and stored at 4°C for no more than 2 weeks before use.

Binding studies

Genes encoding designs were custom ordered from Genscript (Piscataway, NJ) and cloned in-frame into an in-house yeast display plasmid pETCON[2]. For designs targeting HA, Binding studies were done essentially as described[1] using 1 μM of a biotinylated SC/1918/H1 HA1-2 ectodomain. Secondary labels were anti-cmyc FITC (Miltenyi Biotec, Auburn, CA) to monitor design surface expression and SAPE (Invitrogen, Carlsbad, CA) to monitor binding of the biotinylated antigen. For designs targeting *Mtb* ACP2, cells were grown overnight from colonies in SDCAA media and then induced for 24 h at 22°C in SGCAA media. Cells were washed in PBSF buffer (20 mM sodium phosphate, 150 mM NaCl, 1 mg/ml BSA Fraction V pH 7.4) and labeled with 500 nM ACP2-SAPE conjugate for 2 h at 22°C in 1.5 mL eppendorf tubes. Cells were then incubated with anti-cmyc FITC on ice for 10 min. Cells were pelleted at 13,000 xg for 30 s, washed once with 200 μL PBS, and stored as pellets on ice until immediately before reading on a flow cytometer. For designs targeting Fc, yeast cells carrying the design-expression vector were cultivated and induced as described above. They were washed once with PBSM (PBSF with 10 mg/ml BSA) and labeled with 750 nM Rituximab for 4-5 h at 4°C, before adding SAPE (at 1:4 ratio of SAPE to Rituximab) and anti-cmyc FITC antibody (1:100 diluted per volume) for an additional 1h incubation. Cells were washed once with 200 μL ice-cold PBSM and instantly examined *via* flow cytometry.

In all cases, binding signal was quantified as the mean phycoerythrin fluorescence of the displaying population of cells using a 488 nm laser for excitation and a 575 nm band pass filter for emission (appropriately compensated) using either a Cytopeia inFlux Cell Sorter or an Accuri C6 flow cytometer. A binding signal of less than 1.4 of treated to control cells was used as the cutoff for potential binding. Binding studies were repeated at least twice on separate days before designs were discarded. All designs used for this study surface-displayed on the yeast surface. Surface display on yeast requires passage through the endoplasmic reticulum (ER). The ER quality control mechanism restricts some grossly misfolded proteins from reaching the cell surface[4]. Thus it is plausible that a subset of the designed binders used in this benchmark set do not adopt the designed fold.

Per-group scoring methods

Group 1 (Sanbo Qin and Huan-Xiang Zhou, Florida State University)

Our method is based on the electrostatic free energy of the transient complex formed by the protein complex. The transient complex is an intermediate along the pathway to form the native complex[5]. The transient complex separates the bound state, which defines the native complex, with numerous short-range interactions but restricted translational and rotational freedom, from the unbound state, in which the partner proteins form at most a small number of interactions but have expanded translational and rotational freedom. Specifically, the transient complex is located at the outer boundary of the bound-state free-energy well. For many protein pairs, long-range electrostatic attraction has been found to enhance the rates of association[6]. Such cases feature favorable electrostatic free energies in the transient complex, and the rate constants of protein association can be quantitatively predicted[7].

We calculated the electrostatic free energies of ZDOCK benchmark 1.0 set of 59 protein pairs, and found that nearly all of them have favorable electrostatic free energies, both in the native complex and in the transient complex. It therefore seems that electrostatic free energy can be used as a scoring function to select near-native docking poses from non-native ones. In principle, the electrostatic free energy in either the native complex or the transient complex can be used. However, docked poses may contain spurious close contacts between charged groups across the interface. The detrimental effects of such spurious contacts are much reduced in the transient complex, since generally the partner proteins are separated by a layer of solvent in the transient complex[5]. Therefore we settled on using the electrostatic free energy in the transient complex as our scoring function for docking poses.

This scoring function was found to be very successful in this experiment. In the preliminary round (CAPRI Target 43), where 21 complexes were provided, only one of which was a co-crystal structure (see Main Text), we ranked the single native complex at the very top. In the comprehensive design discrimination benchmark (CAPRI Target 44), we ranked the only complex with evidence of binding (design 45)

as fourth, with very small differences in electrostatic free energy from the top three ranked models.

We followed the same method for Target 45. For each model system, 50 steps of steepest-descent energy minimization were performed on hydrogen atoms to remove potential clashes. The ensemble of configurations representing the transient complex for each model system was generated using a previously developed procedure[5]. Briefly, the ligand was translated and rotated around the putative native complex, as given to us in Target 45. The rotation around the axis perpendicular to the least-squares plane of the interface exhibits a characteristic sharp transition: the range of allowed rotation angles is very limited in the bound state but rapidly widens as the transient complex is passed. This sharp transition allowed the transient complex to be identified.

We then calculated the electrostatic free energies on 10 representative configurations of the transient complex and used their average as the scoring function. The electrostatic free energies were calculated by the APBS program (version 1.2.1), with the following parameters: the grid dimensions were 193 × 193 × 193, with coarse grid size at 1.5 Å and fine grid size at 0.5 Å. The dielectric boundary was defined as the van der Waals surface of the solute molecule[5] (set with the option "srfm mol, srad 0.00"). The atomic partial charges were those of the AMBER force field and were distributed to the grid with the option "chgm spl2". The temperature was set to 298 K. The solute and solvent dielectric constants were 4 and 78.5, respectively. The ionic strength was 60mM.

Finally the average electrostatic free energy of the transient complex was transformed to a normalized score, using the following scheme: < -1 kcal/mol in electrostatic free energy corresponds to a score of 1 (binds); between -1 and 0 kcal/mol a score of 2 (likely to bind); between 0 and 1 kcal/mol a score of 3 (likely not to bind); between 1 and 2 kcal/mol a score of 4 (does not bind); and higher electrostatic free energies correspond to a score of 5 (uncertain).

Group 2 (J.C. Mitchell and O.N.A Demerdash, University of Wisconsin, USA)

Our model combines biophysics and informatics. It is very efficient, requiring only seconds to calculate. To create the model, we applied support vector machines to a large number of energetic features calculated for each protein interface. Our goal was to classify a given structure as a binder or non-binder and provide a score able to distinguish between the two classes for most examples.

A data set was produced using four different types of binders and two different types of nonbinders. The nonbinders used for training included docking decoys as well as a subset of the nonbinders that were part of the CAPRI 20 prediction exercise. Examples of binders included crystal structures from three different data sets and near-native docking predictions. In each case, a relatively small amount of the data was used for training and the remainder used for testing. We believe the use of diverse training data helps make the model more robust, and this is clearly indicated by its performance.

Our complete feature set consisted of a large number of energetic terms that were implemented for use with our docking program, *ReplicOpter*[8]. These terms include six electrostatics potentials, four hydrogen bonding potentials, stacking/pi interaction potentials, a softened van der Waals potential, atomic contact energy, and shape specificity.

Using different small training sets created from the data described above, we could obtain many models with the following characteristics:

1. classifies crystal structures for the nontraining test sets with at least 75% accuracy
2. classifies Rosetta-designed (presumed) nonbinders with at least 80% accuracy
3. scores Rosetta-relaxed complexes similarly to the original crystal structures
4. strongly classifies the Rosetta-designed binder for which binding is confirmed (#10 from Target 43)

Complete details on the training procedure, along with references and descriptions of all energetic terms and data sets, will be provided in a separate manuscript.

Group 5 (Mayuko Takeda-Shitaka and Genki Terashi, Kitasato University, JP)

In order to provide a valid discrimination between the interfaces that bind and those that do not bind, we constructed a scoring function from the training data set (1619 protein-protein interfaces, NR70% of sequence identities). In this experiment, we used a newly developed atom-atom potential instead of a previously developed residue based scoring function (CIRCLE QA program [9]) used in SKE-DOCK [10].
The pairwise atomic potential between the atom of type *i* and *j* at the interface can be described as:

$$AA_{i,j} = \frac{1}{N_{training}} \sum_{m}^{N_{training}} \left(\log \frac{1.00 + w \cdot f_{native}(i, j, m)}{1.00 + w \cdot f_{decoy}(i, j, m)} \right)$$

where $N_{training}$ is a number of training data set (=1619), $f_{native}(i,j,m)$ is a frequency of *i-j* contacts occurring in protein-protein interactions of the *m*th native structure, $f_{decoy}(i,j,m)$ is that of 100 decoy models, which are obtained from the FFT based rigid-body docking method (like ZDOCK[11]). In the usual case, the native structures were unknown. Therefore, the near-native solutions were not filtered from the decoy models. The optimized cutoff distance of atom-atom interaction and weight (*w*) are 6.0Å and 0.02, respectively.

For scoring, the given 207 structures were re-docked by rigid-body docking. Then according to the shape complementarity, best 100 models were selected as decoy for each structure. The pairwise atomic potential were summed up and Z-scores were calculated by comparing with potential distributions of 100 decoy models. A high Z-score means that the protein-protein interface of the given structure has better pairwise atomic potential

than decoys. We assigned the 207 structures into five classes according to the Z-scores as follows:

1 (binds) 3.0 <= Z-score

2 (likely to bind) 2.0 <= Z-score < 3.0

3 (likely not to bind) 1.0 <= Z-score < 2.0

4 (does not bind) Z-score < 1.0

5 (uncertain) could not finish the re-docking step due to the technical problems

Group 6 (Iain H. Moal, Xiaofan Li and Paul A. Bates
Cancer Research UK London Research Institute, UK.)

All 207 structures were redocked, globally with SwarmDock[12], and locally with PyRosetta[13]. Encounter complex formation and dissociation was simulated with BioSimz[14]. A number of metrics, which characterise the results of these computations, were derived. Numerous interface descriptors and binding energy terms were also calculated: the analytical continuum solvent (ACE) potential terms[15], DComplex[16], Rosetta energy terms[13], interface packing and surface complementarity scores[17], and generalised Born (GBSW) electrostatic and non-polar solvation energy[18]. Further parameters, describing interface flexibility and flexibility differences of core and peripheral interface residues, were calculated using elastic-network normal-mode analysis[18], as well as counts of the number of residues with binding energies below various energy thresholds and the number of buried hydrogen-bond (H-bond) donors and acceptors. The distribution of these parameter values indicates that, compared to the designed, the benchmark complexes have a lower H-bond binding energy, fewer unsatisfied buried H-bond donors and acceptors, a more favorable change in ACE energy upon binding, and more favorable encounter-complex formation dynamics. These parameters are sufficient to distinguish the benchmark and designed complexes, as a parameter set capable of binomial classification at 97.6% accuracy (95.7% accuracy with leave-one-out cross validation) was found, using a support vector machine with an analysis of variance kernel and a population-based forward greedy feature selection algorithm. However, as the

designed complexes are not to be used in the discrimination of the two categories, the parameters were used in linear combination, in an energy function trained on empirically derived binding free-energy values.

Dissociation constants for 95 complexes in the Benchmark 3.0 were manually amalgamated from the literature and empirical binding free energies were calculated. This affinity benchmark was later expanded upon[19] and is available on-line (http://bmm.cancerresearchuk.org/~bmmadmin/Affinity/). Linear regression of a number of parameter sets were performed, for which thresholds could be found to discriminate the benchmark complexes from the designed with between 70% and 90% accuracy. Values shown correspond to an energy function composed of the following parameters, along with their relative importance (derived from normalized weights), where the first two parameters are Boolean: Is SwarmDock top ranked structure under 5Å root-mean square deviation (RMSD) to bound? (0.076), does the biggest SwarmDock cluster correspond to the bound? (0.009), Rosetta all-atom pair potential (fa_pair; 0.299), Rosetta coarse-grained pair potential (0.122), van der Waals (0.192), BioSimz predicted kon constant (0.027), interface packing (0.181), surface complementarity (0.176), ACE self-solvation energy (0.780) and GBSW solvation energy (0.071). The regression has a RMS error of 2.76 and a correlation of 0.414, has an area-under-ROC of 91.2% and can be used to correctly classify 88.4% of the complexes with a threshold of -9.55kcal/mol. Bins were chosen such that the 'binds' category contains complexes with predicted binding energies below -9.8 kcal/mol, and the cutoffs for the higher bins, likely to bind, uncertain, likely not to bind and don't bind, are chosen at intervals of 0.2 kcal/mol.

Group 7 (Martin Zacharias)

A composite score based on three equally weighted components was used to evaluate protein-protein complexes. It consists of the docking score after docking re-minimization in rotational and translational coordinates of protein partners employing the program ATTRACT[20] and a coarse-grained force field for the protein partners[21]. The starting structure for docking minimization was the provided model or native complex structure (hereafter termed reference structure). The scoring energy unit is 1 RT, where R is the gas constant; and T is room temperature).

The ATTRACT score is a knowledge based docking scoring function optimized on a large number of protein-protein complexes[21]. Typically, the absolute value of the ATTRACT score for minimized complexes gives an impression if a complex is favorable or not. However, occasionally the score alone is insufficient.

The second contribution to the composite score is related to the deviation of the minimized complex from the starting structure (RMSD of the ligand multiplied with 1 RT/Ångstrom).

This contribution is based on the observation that for experimental structures of protein-protein complexes docking minimization using ATTRACT results in a minimized complex with an RMSD (of the ligand relative to the fixed receptor protein) of < 1-3 Å. Deviations larger than 3-5 Å may indicate unrealistic complex structures.

The third contribution to the composite score required a systematic docking run over the whole surface of the protein partners [following published protocols, ref. [20]]. If the rank of the docking minimum closest to the reference structure was within the 10 top ranked solutions a scoring penalty of 0 RT, for rank 10-100 a penalty of 6 RT, for rank 100-1000 a penalty of 12 RT and for rank > 1000 a penalty of 18 RT was added, respectively. The rationale for this contribution is:

For experimental complex structures (bound partners) a systematic ATTRACT docking search typically gives the docking minimum closest to experiment as top ranking solution or within the 10 best scoring solutions. If this is not the case for a given protein-protein complex it indicates that the complex structure is probably unrealistic.

The above composite score was transformed to a normalized score:

1. raw composite score < -6 RT => normalized score 1 (binds).
2. -6 RT < raw score < -3 RT: normalized score 2 (likely to bind).
3. -3 RT < raw score < 0 RT: normalized score 3 (likely not to bind)
4. raw score > 0 R : normalized score: 4 (does not bind).

normalized score 5 (uncertain) was not needed.

For the largest fraction of the Zdock structures the raw score was better -6 RT

(rationale for choosing -6 RT as limit for normalized score 1). For very few Zdock-cases a raw score above 0 RT was found. Therefore. 0 RT was chosen as the limit to distinguish non-binders (score 4). With this scheme 108 of the 118 Zdock complexes scored 1 or 2 and 60 out 87 of the designed complexes scored 3 or 4 (12 scored 1).

Group 8 (Hahnbeom Park, Jun-su Ko, Hasup Lee, and Chaok Seok, Seoul National University, Korea)

A discrimination score expressed as a weighted sum of the following seven terms was developed: 1) the DFIRE potential[22], 2) van der Waals energy with CHARMM19 parameters, 3) Coulomb energy with a distance-dependent dielectric constant and CHARMM19 parameters, 4) an empirical solvation term described by solvent-accessible surface area and atomic solvation parameters (ASP)[23], 5) a knowledge-based orientation-dependent hydrogen bond energy[24], 6) a sequence-conservation score derived from PSI-BLAST profile[25], and 7) a sidechain entropy derived from iteratively calculated probability distribution of interface rotamers[26].

The seven weight parameters for the discrimination score were obtained by minimizing the number of complexes in the overlapping score region where binding and unbinding training complexes coexist. The training set for binding complexes contains 74 binding complexes with known binding affinities[27] and that for unbinding designed complexes are those provided by the Baker group as CAPRI round 20 targets that were found not to bind. Contribution of each term to discrimination was assessed by means of normalized weights. The sequence-conservation score contributes the most to the total score (52%), and the hydrogen bond energy (18%) and the sidechain entropy (13%) follow. Contributions by DFIRE and solvation term are almost negligible.

The score regions for the five categories were determined based on the distributions of the binding and unbinding complexes as follows: binds (< -8.0), likely to bind (-8.0 ~ -6.0), uncertain (-6.0 ~ -1.0), likely not to bind (-1.0 to 0.0), and does not bind (>0.0).

Group 9 (Thomas Bourquard, Julie Bernauer, Anne Poupon, Jérôme Azé, INRIA AMIB / INRA Tours, France)

Our method relies on the use of machine learning to build a scoring function able to discriminate between native or "near-native" structures and decoys. The native structure set contains 211 native complex structures. The set is made of 187 bound-unbound and unbound-unbound complexes described in a previous study[28] updated with newly determined structures.

Non-native structures (or decoys) used to train the procedure were generated with an in-lab geometric generation procedure. Like in our previous work[29-31], a coarse-grained model with one point per residue, called node, is used. For each complex, the Delaunay triangulation of the partners' nodes is computed using CGAL (http://www.cgal.org) and its dual, the Voronoi tessellation is build. A pseudo-normal vector is then built for each node by summing the vectors corresponding to the neighboring edges in the triangulation oriented towards the solvent and having a fixed length of 6.5 Å. For each possible pair of vectors (one in each partner), the ligand partner is translated and rotated to bring the two vectors point to point and in opposition. The ligand is then rotated around this axis and a conformation is built every 5Å. This method does provide near-native solutions for all tested complexes. For each native structure, decoys having an interface area larger than 400Å² for which enough parameter categories were represented and having a RMSD larger than 10 Å relatively to the native structure were kept.

96 training attributes were considered based on the properties of the residues and pairs present at the interface. For pair attributes, residues are binned in six categories: hydrophobic (ILVM), aromatics (FYW), small (AGSTCP), polar (NQ), positively (HKR) and negatively charged (DE). The attributes are: the Voronoi interface area, the number of residues at the interface, the fraction of each residue type at the interface, the mean Voronoi volume of the interface residues, the fraction if each pair type at the interface, the mean interface

node-node distance, the fraction of interface residues for each category and the mean Voronoi volume of interface residues for each category.

Learning is performed using a genetic algorithm optimizing the area under the ROC (Receiver Operation Characteristics) curve. We used a ?+µ scheme, with 10 parents, 80 children and 500 generations and classical cross-over and auto-adaptative mutations. The scoring functions are expressed as S(conf)=wi|xi(conf)-ci| where for each attribute xi, wi and ci are the weight and centering value respectively. These attributes are optimized through the learning procedure. All functions were learned in a 10-fold cross-validation setting.

The strategy has proven effective in previous work[29; 31] and in the previous CAPRI rounds. To evaluate whether our scoring function can be used to predict association, we evaluated its performance relatively to the binding affinity benchmark published in the study by Kastritis et al.[32]. In its raw version, the scoring function performs slightly better than PISA. We decided to use this strategy in its raw form for the Rosetta decoys (no special learning was made). As no clear signal was observed to discriminate between the binding categories offered, we defined them according to the score distribution obtained on the ZDOCK 3.0 benchmark. A score above the 3rd quartile is considered representative of the first category “1-binds”. A score above the median but below the 3rd quartile was labeled “2-likely to bind”. A score between the median and the mean was considered “5-uncertain”. A score between the 1st quartile and the mean was labeled “3-likely not to bind”. A score below the first quartile was labeled “4-does not bind”. On average Rosetta decoys scores show that they are less likely to bind than the ZDOCK 3.0 benchmark examples.

Group 10 (Seren Soner[1], Sefik Kerem Ovali[1], Pemra Özbek[1], Nir Ben Tal[2], Türkan Haliloglu [1], [1]Polymer Research Center and Chemical Engineering Department, Bogazici University, Bebek - Istanbul, Turkey, [2]Department of Biochemistry and Molecular Biology, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel)

We used the Anisotropic Network Model (ANM)[33; 34] to predict the residue fluctuations and collectivity of the motion[35] in the bound versus unbound states of the chains in the complex structures of the ZDock Benchmark and Design datasets. The premise is that the global mode of motion of a complex structure results from the collective participation of substructures or chains, where the chain's degree collectivity could be expected to increase upon formation of the biological complexes in the Benchmark set but not in the Design set.

In the ANM[33], where the protein structure is modeled as an elastic network, the correlation between the residue fluctuations $\Delta\mathbf{R}_i$ and $\Delta\mathbf{R}_j$ of residues i and j is calculated as $\left\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \right\rangle = \frac{(3k_BT)}{g} tr\left[\mathbf{H}^{-1}\right] = \frac{(3k_BT)}{g} \sum_k tr\left[l_k^{-1}\mathbf{u}_k\mathbf{u}_k^T\right]$

Here $\mathbf{H}^{-1}$ is the inverse of the Hessian matrix. $\mathbf{u}_k$ and **?**$_k$ refer to the eigenvectors and eigenvalues of 3N-6 modes, respectively, N being the number of residues. The degree collectivity (K) of a motion by any individual mode k is defined as proportional to the exponential of "the information entropy" of the eigenvector k[35].

$$K_k = \frac{1}{N} exp\left(-\sum_i^N a\Delta\mathbf{R}_i^2 log\left(a\Delta\mathbf{R}_i^2\right)\right)_k$$

The normalization factor a is chosen such that $\sum_i^N a\Delta\mathbf{R}_i^2$ equals one. In the most collective motion all ? R_i^2 are expected to be identical, and K approaches 1, whereas in the extreme local motion K approaches zero for a large chain (K = 1/N). Here we considered only the fluctuations in the slowest mode to estimate the collectivity of the motion of the chains in the bound and unbound states. The unbound state was taken as the co-crystal structure of the complex.

The increase in degree collectivity of chains A and B upon complex formation and the difference in the increase in degree collectivity between chains A and B, DiffColl, was calculated for both Benchmark and Design datasets. The increase in chain A's collectivity upon binding was noticeably lower in the Design set than in the Benchmark set, whereas chain B's collectivity was higher in the Design set than in the Benchmark set. Further, the collectivity increase between the two chains of a given complex structure was more similar in magnitude in the Benchmark set than in the Design set. DiffColl, which is the difference between the increase in the degree collectivity of chains A and B, appears as a plausible measure to identify the biological complexes. A success rate of 72 % and 71 % was obtained for the Benchmark and Design datasets, respectively, with the following criteria: Binds if DiffColl > 0; Likely to bind if -0.24 < DiffColl < 0; Likely not to bind if -0.35 < DiffColl < -0.24; Does not bind if DiffColl < -0.35.

Overall the results suggest that the global dynamics is a major discriminant of biological association. It is very encouraging that a single criterion, based on first principles, is useful for discrimination between true complexes and artefacts. Hopefully, it could be integrated with measures of other qualities of protein complexes to improve the overall performance.

Group 11 (Howook Hwang, Thom Vreven, Brian G. Pierce, Zhiping Weng, University of Massachusetts Medical School, Worcester, MA, USA)

We applied the ZRANK scoring functions for re-ranking developed in our laboratory. This includes the original version of ZRANK [36], ZRANK extended and reparameterized to score structures that are refined using Rosetta (ZRANK 2.0) [37], as well as the recent function that combines atomic and residue based terms (IRAD) (in preparation). The ZRANK functions are linear combinations of weighted terms, and we did not optimize scoring functions for this specific goal, although the datasets that were used to determine some of the components did include cases from the Benchmark. ZRANK requires the

structures to include hydrogens, which were added using the Rosetta package[38].

Based on the area under the ROC curve (AUC, with cases from the Benchmark as positives, and the designed cases as negatives) of the ROC curve (with cases from the Benchmark as positives, and the designed cases as negatives), the original ZRANK function performs the best of the three functions, although the differences are not large. This function was therefore selected for comparison with the methods from other labs. To better understand the performance of ZRANK, we also calculated the AUC's of the individual terms of ZRANK, as well as the count of atom contacts between the binding partners, with a 6 Å cutoff, as a measure of interface size. We found that the ZRANK terms that strongly correlate with the interface size (attractive van der Waals, attractive electrostatic interaction, and the 6 Å binding partner contact count) all perform as well or better than the complete ZRANK function. It is clear that the main discriminating feature between the positives and negatives is the size of the interface. The interface sizes of the designed complexes are smaller than typically observed in the Benchmark, which may be related to the design process itself.

Because the set of complexes contains 120 cases known to bind, the top 60 scoring were assigned a 'binding' normalized score, and the next 60 'likely to bind'. The remaining 87 predictions were distributed in three sets of 29 predictions over the 'uncertain', 'likely not to bind', and 'does not bind' categories.

Group 12 (Laura Pérez-Cano, Carles Pons, Juan Fernández-Recio, Barcelona Supercomputing Center, Spain)

We have checked our standard pyDock scoring function[39] on the set of cases with experimentally determined protein-protein binding affinity collected by Kastritis and Bonvin[32], We used the 77 complex structures that were minimized with Rosetta (to use the same minimization protocol as the provided non-binders). Although the correlation between pyDock energy and the experimental values on the global set was quite low (r = 0.21), when we considered only those cases in which the experimental energy was obtained

by ITC studies, the correlation was much larger (r = 0.76). Interestingly, we observed that the individual energy terms in pyDock scoring function (*pyDock energy* = *electrostatics* + *desolvation* + *van der Waals*) showed different correlation levels with respect to the experimental data (r = 0.69 for *electrostatics*; r = -0.27 for *desolvation*; r = -0.04 for *van der Waals*). This suggested that although our scoring function was optimized for the identification of near-native orientations in rigid-body docking, it might be still improvable for the prediction of binding affinities. Given the insignificant correlation of the van der Waals term with the experimental values, and the positive and negative correlation of the electrostatics and desolvation terms, respectively, we defined a binding affinity predictor function *pyDockAFF* as follows, without any further optimization of weighting factors to avoid over-training:

pyDockAFF = *electrostatics* - *desolvation* (1)

We then established confidence thresholds for *pyDockAFF* based on the capability to discriminate strong from weak binders in the 77 cases compiled by Kastritis and Bonvin [2] and minimized with Rosetta. For that, we arbitrarily defined strong or weak binders as those with experimental binding energy smaller or higher than -8 kcal/mol (i.e. μM affinity), respectively. We observed that the median value of *pyDockAFF* for the strong binders was around -15 kcal/mol (that is, half of the strong binders had better *pyDockAFF* value), so we decided to use this cutoff to classify the cases as "binders". Similarly, we observed that the median value of *pyDockAFF* for the weak binders was around 0.0 kcal/mol, so this value could be used to separate binders from non-binders. Thus, we defined as "uncertain" those with *pyDockAFF* values between -5 and +5 kcal/mol. As for the other categories, we just chose the values that generated same sized intervals. The summary of categories (and the corresponding score numbers in our submission) is the following: "binds" (score 1): *pyDockAFF* < -15.0; "likely to bind" (score 2): -15.0 < *pyDockAFF* < -5.0; "uncertain" (score 5): -5.0 < *pyDockAFF* < +5.0; "likely not to bind" (score 3): +5.0 < *pyDockAFF* < +15.0; "does not bind" (score 4): *pyDockAFF* > +15.0.

Group 14 (F. Jiang and co-workers, Institute of Physics, CAS, China)

First, 120 bound structures from the ZDock Benchmark 3.0[40] as provided in the Rosetta-relaxed conformations[38] were used to develop a set of linear regression weights to calculate a combined score. This was done by a simulated annealing program IntfacMove, which allows sampling of different rigid body configurations and side chain rotamers. 2000 steps were simulated for each complex to generate near-native conformations. The potential energy terms included were (1) van der Waals interaction, for which 6-8 potential was used. (2) electrostatic interaction, which is Coulomb potential with a distance-dependent dielectric constant. (3) solvation effect, which is calculated by counting the number of atoms buried by the interface according to atom types, for which DeLisi[41] 18 atom types were used. (4) atomic contact potential as defined by DFIRE[22]. (5) atomic contact matrix with three element types, namely, polar-polar, nonpolar-nonpolar and polar-nonpolar. (6) hydrogen bond matrix with three element types, namely, acceptor to donor, acceptor to acceptor or donor to donor, and acceptor or donor to other atom types. (7) center-of-mass attraction between the ligand and the receptor, implemented as a harmonic potential. (8) atomic pair contact repulsion at the interface, implemented as a harmonic potential. (9) many-body interaction between nonpolar atoms; (10) many-body interaction between charged atoms; (11) many-body interaction between hydrogen bond donors and acceptors. Many-body interaction for 4-body (node) graphs is calculated by counting the number of different graphs with different topologies. Many-body interaction for 5, 6, and 7-body graphs is calculated by counting graphs with different number of interaction edges. Only up to 7-body interaction graphs are considered. These potential energy terms were used as the free variables to fit the experimental value in the linear regression. The experimental value is fnat, the fraction of native atomic contacts relative to the input starting structure of the simulation by IntfacMove. From the merged results of 120 simulations, a set of overall weights was obtained, independent of the 87 designed structures and other decoys. The correlation coefficient of the linear regression fit is 0.794. Inclusion of many-body interaction seems to be significant in the final scoring, although its effect on the correlation coefficient

is moderate, 0.76 if not including the many-body interaction terms in the potential energy.

Then, both the 87 designed and 120 Benchmark structures were calculated for the combined fnat-like scores. These scores were then converted to z-scores using an average and standard deviation as background noise calculated by generating 80 decoy structures provided by ZDOCK3[40] and simulating with the same procedure IntfacMove as for the 120 Benchmark structures as described in above. The classification of binding is determined by z-score: in the range of [-inf, -3]=bind; (-3, -2]=likely to bind; (-2, -1]=uncertain; (-1, -0.5]=likely not to bind; (-0.5, +inf)=do not bind. Out of the 120 Benchmark structures, only two were in the wrong category: one (2PCC) was in category 4 and one (2OOB) in category 3. For the 87 designed structures, two were in category 3 and none in category 4. Most of them, 64.4 %, were in category 5, while for the Benchmark structures, the percentage was 30%. For category 1, the percentage of the Benchmark and the designed were 40% and 3%, respectively, while for category 2, they were 28% and 30%.

Group 16 (Feng Yang, Xinqi Gong, Libin Cao, Xianjin Xu, Bin Liu, Panwen Wang, Chunhua Li, Cunxin Wang. College of Life Science and Bioengineering, Beijing University of Technology, 100124, China.)

In order to discriminate the native interfaces from non-native ones, we first tried to analyze all the complexes qualitatively using two methods. In one method, the interaction patches across interfaces were analyzed using our prediction method for the binding site patch[42]. The complexes whose binding sites are not consistent with the predicted ones were classified as non-binders. In the other method, we calculated interface areas and defined those structures with interface areas less than 1200 $Å^2$ as the non-native ones.

But both methods mentioned above cannot produce quantitative scores for every structure. Therefore, we used two scoring functions, HPNCscore[43] and rpscore[44], to evaluate all the decoys. To test the discriminative power of these two functions, we applied them on the constructed database containing 120 correct and 120 wrong protein-protein complex structures. From the distribution results of the correct and

wrong structures, we found that rpscore has the better prediction ability and hence we used the rpscore values as the raw scores of decoys.

According to the rpscore values of 207 structures combined with the qualitative analyses mentioned above, we empirically defined the five normalized categories as following: the models with raw value = -13.70 are categorized as "does not bind", -16.50 = raw value < -13.70 are categorized as "likely not to bind", -20.00 = raw value < -16.50 are categorized as "uncertain", -21.05 = raw value < -20.00 are categorized as "likely to bind", raw value < -21.05 are categorized as "bind".

Group 17 (Charles H. Robert and Mainak Guharoy, Laboratoire de Biochimie Théorique CNRS-UPR 9080, Institut de Biologie Physico-Chimique, Paris, FRANCE).

Our model uses evolutionary information to score protein complexes. We had previously shown that the interface core is more conserved than the rim region in most biologically relevant complexes. A dataset of crystal contacts served as a negative control in that study, showing that no such distinction of core versus rim was seen in non-specific interfaces[45].

For the current challenge of discriminating 'binders' from 'non-binders', the ZDock benchmark of native protein complexes was used as binders for training. The non-binders consisted of docking decoys obtained using ZDock to dock the paired components. None of the CAPRI 21 designs were used in the training.

Evolutionary information for each protein was obtained in the form of multiple sequence alignments from the HSSP database[46], which for each PDB entry provides a multiple sequence alignment for structurally homologous proteins. For the designed complexes, we first performed a BLAST[47] search to identify the closest homolog in the PDB in order to obtain a representative alignment. Using the multiple alignment, sequence entropy values were calculated for each individual interface residue (both in the core and rim). Interface residues were identified based on the change in solvent ASA upon separating the components in the complex – any residue gaining more than 0.1Å^2 ASA was considered to belong to the interface[48]. Mean entropy values <s> were then calculated separately for the core and rim regions after weighting the individual entropy values by the corresponding residue DASA, and the ratio

$\langle s \rangle_{core}/\langle s \rangle_{rim}$ is obtained. More detail can be found in[49]. For biological interfaces $\langle s \rangle_{core}/\langle s \rangle_{rim}$ is generally less than 1.0, implying that the core is more conserved on average than the rim[49].

Scoring was performed by a simple binary method. Mean core/rim entropy ratios were calculated both for the set of native ZDock complexes and for the set of decoy structures. The average of these two means was taken as a threshold value – an entropy ratio lower than the threshold was taken as indicating a 'binder' (we assigned these cases to the category 'likely to bind'), whereas an entropy ratio greater than the threshold was taken as indicating a 'non-binder' (assigned to the category 'likely not to bind'). Using this simple scheme, 54% (47/87) of the designs were scored "likely to bind", versus 77% (95/124) of the Rosetta-minimized native complexes.

Group 20 (Shiyong Liu, Yangyu Huang, Lin Li , Dachuan Guo, Ying Chen, Yi Xiao, Huazhong University of Science and Technology, China).

Our scoring function DECK-X here is evolved from a knowledge-based potential DECK[50]. In DECK, each residue is represented as one pseudo-atom, the centroid of the side chain(SCM). In order to capture more interactions between residues, two points (CA and SCM) are used in our current scoring function DECK-X.

$DECK\text{-}X = w_1 * E_{11} + w_2 * E_{12} + w_3 * E_{21} + w_4 * E_{22} + w_5 * N_{clash}$

$$E_{mn} = \sum_i \sum_j e_{ij}(mn)$$

Where E_{mn} (m=1, 2; n=1, 2) is the interaction energy between CA (m=1) or SCM (m=2) of residue i and CA (n=1) or SCM (n=2) of residue j from receptor and ligand, respectively. N_{clash} is the number of clashes of the protein-protein interface. $e_{ij}(mn)$ is derived from the following equation:

$$e_{ij}(mn) = -RT \ln \frac{p(i, j, d(m, n))}{p^*(i, j, d(m, n))}$$

p(i, j, d(m, n)) is calculated from near-native decoys (ligand RMSD less than 5 Å). Our reference state p*(i, j, d(m, n)) is calculated from non-near native decoys (similar to RAPDF[51], PIPER[52], DARS[53]).

In order to optimize 5 parameter, our target function is adapted from the term G1[54] in the following:

$$G_1 = \frac{1}{1+\frac{1}{351}\sum_{k=1}^{351}\frac{\left\langle R(k,j)\sum_{i=1}^{Np} w_i E_i(k,j)\right\rangle_j - \left\langle \sum_{i=1}^{Np} w_i E_i(k,j)\right\rangle_j \left\langle R(k,j)\right\rangle_j}{\left(\left(\left\langle (\sum_{i=1}^{Np} w_i E_i(k,j))^2\right\rangle_j - \left\langle \sum_{i=1}^{Np} w_i E_i(k,j)\right\rangle_j^2\right)\left(\left\langle R(k,j)^2\right\rangle_j - \left\langle R(k,j)\right\rangle_j^2\right)\right)^{1/2}}}$$

Where R(k,j) is the LRMSD of jth decoy structure of kth training protein-protein complex. N_p (=5) is the number of undetermined parameters (w_i values) of DECK-X. $E_i(k,j)$ is the scoring term related to the parameter w_i. $\langle\cdots\rangle_j = (1/10000)\sum_{j=1}^{10000}\cdots$

We tried to maximize the correlation coefficient between the LRMSD and the total binding energy for 3487680 GRAMM-X[55] docking decoys of 351 protein-protein complexes. The L-BFGS method (http://chokkan.org/software/liblbfgs) is used to solve the minimization problem. Finally, we got a set of parameters: {w1, w2, w3, w4, w5} = {-4.654474, 2.209604, 2.561927, 3.677414, 0.733294 }. Then, the CAPRI Scoring target was tested by our DECK-X directly. Finally the DECK-X score was translated to a normalized score as following:

Normalized score is set as 1 when -785 < DECK-X < -401; it is set as 2 when -386 < DECK-X < -308; it is set as 3 when -296 < DECK-X < -202; it is set as 5 when -198 < DECK-X < -100; it is set as 4 when -100 < DECK-X < -22.

These bins are selected arbitrarily according to experimental information: most designed protein-protein complexes do not show binding at all.

Group 21 (Nir London, Zohar Itzhaki, Ora Schueler-Furman, Department of Microbiology and Molecular Genetics, Institute for Medical Research Israel-Canada, Hadassah Medical School, The Hebrew University, POB 12272, Jerusalem, 91120 Israel.)

In order to select a metric that would be able to discriminate between the native interfaces of the benchmark 3.0 complexes to the models of non-binding designs, we assessed different realistic interface descriptors on the entire set of interfaces. The discrimination ability of each descriptor was measured as the area under the curve (AUC) for a receiver operator characteristic (ROC) plot, in which the ZDock Benchmark 3.0 complexes were

considered as 'binders' and the set of interface designs were considered as 'non-binders'. This is similar to the approach we described for FunHunt - an algorithm to discriminate between true and false binding funnels, and a description of the different parameters can be found therein [56].

We list the evaluated features, in decreasing discrimination ability (AUC in parentheses): Polar solvent accessible surface area (SASA) buried at the interface (0.85) - Designs display smaller buried polar SASA; Interface solvation energy (0.85) - Designs display better solvation energy as evaluated by Rosetta; Total SASA buried at the interface (0.82) - Designs display an overall smaller interface size; Interface attractive Van-der-Waals (VDW) term (0.78) - Native complexes display somewhat better 'attractive' values; Interface repulsive VDW term (0.76) - Native complexes display somewhat worse 'repulsive' values; Apolar solvent accessible surface area (SASA) buried at the interface (0.76) - Designs display smaller buried a-polar S ASA; Unsatisfied backbone hydrogen bond donors/ acceptors buried at the interface (0.69) - Native interfaces display somewhat less such unsatisfied groups; Interface hydrogen-bond energy (0.68) - Native interfaces display better hydrogen bonding energy; Rosetta interface DDG (0.66) - Native interfaces display better DDG's; Statistical pair potential (0.58) - Native interfaces show slightly better pair potential.

We chose the metric of polar solvent accessible surface area buried at the interface as a discriminator between native complexes and designed complexes, resulting in the highest AUC discrimination value. While other terms that show similar discrimination might indicate problems in the Rosetta energy function (such as the solvation term which shows better values for designs), this measure shows the best discrimination between the benchmark 3.0 complexes and the designed interface and is based on a realistic physical measure.

Calculation of polar solvent accessible surface area buried at the interface: For a protein complex AB, the measure of polar solvent accessible surface area (SASA) buried at the interface of proteins A & B is calculated by taking

the difference between the polar SASA of the complex AB and the individual free partners: Buried polar SASA = Polar SASA(AB) - polar SASA(A) - polar SASA(B). The polar SASA of a given protein is calculated by rolling a probe of radius 1.4A over the surface of the protein. The fractional exposed surface area of each atom is calculated based on ref. [57]. Each atom is categorized as polar or apolar, and the polar SASA is the summation of the fractional exposure of all polar atoms.

Group 22 (Gideon Schreiber, Yuval Inbar, Mati Cohen, Vladimir Potapov)

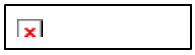
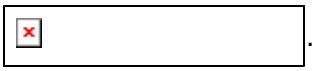
A number of supervised learning techniques utilizing different scoring functions were used to distinguish between interacting and non-interacting complexes. Any supervised learning technique relies on an input vector, in our case various scoring features of protein complexes, and an answer vector, in our case the classification of the complexes. Since we lack the binding energies for some native and for all decoy complexes, we decided to classify them only into binding and non-binding rather than 5 different classes. Moreover, we have focused on minimizing the false binders rate, as there are many more false than true complexes in a realistic scenario.

Training set: In order to distinguish between interacting and non-interacting proteins, we constructed native and decoy data sets. The native data set was the *ZDOCK* sets number 2 and 3[40]. To introduce a variation in the decoy set we joined two different types of non-binding complexes. The first is a collection of crystal contact complexes[58]. The second is a collection of mis-predicted docking solutions of the *Rosetta* group as submitted in a previous CAPRI round[59]. Only submissions with more than 4Å RMSD from the native complexes were included. All complexes, negative as well as positive ones, underwent rotameric minimization using the Hunter energy function.

Scoring functions**:** 24 different scoring features of various interface properties were considered. These include; interface area, electrostatic energy of interaction (calculated by PARE[60]), side chain interaction (evaluated by Hunter), rotameric probabilities, solvation and Lennard-Jones (for more details see [61]), as well as residue contact map score (see below for details) and geometric score.

Contact Map Score: An accumulative residue contact map represents the number of contacts between the residues type pairs. We define a contact by the minimum distance between sidechain atoms (if it is less than the sum of VDW radii + 1.8 Å). Given the residue interface composition of both proteins, we can compute an expected contact map based on statistics that were extracted from contact maps of solved structures. We score a given contact map based on the difference between the contact values of the given and expected contact map, as defined by equation (1) for *contact map score 1*:

(1) 

Where X is a 20x20 matrix that represents the accumulative residues contact map (x_{ij} is the number of contacting residue pairs of type i and j, where the residues of type i and j belong to molecules 1 and 2 respectively). I^1 and I^2 are the interface composition vectors of molecules 1 and 2 (I_k is the number of residues of type k in the interface).  is the expected number of contacts between the residue types i and j, and  is its standard deviation. We compute the expected value and the standard deviation by assuming a binomial distribution of the contact number over the number of potential contact pairs  and the probability p_{ij} of a pair to be in contact given they are in the interface. Hence,  and  .

The values for the different p_{ij} 's were computed using the observed contact maps of 620 non-redundant native complexes (30% identity or less).

Contact map score 2 is also based on the difference between expected and observed values, however it favors contacts between frequently contacting residues.

Learning: We tested three different supervised learning techniques 1.Decision tree learning (MathLab); 2.Linear classifier (an in-house algorithm) and 3. ***LIBSVM*** a support vector machine (SVM) algorithm.

The best results were achieved using SVM algorithm limiting the combinations to up to 16 different scoring features (preferable less). The optimal combination consisted of 7 features: *PARE*, *Hunter*, surface area, non-polar

interface area, knowledge based *contact scores 1 & 2* and the number of interacting pairs. The SVM correctly classified 56 out of the 120 native complexes and, more importantly, 216 out of 224 decoys (only 8 false positive). When applied on the designed set it has wrongly classified 4 as positive (probably false) out of 87. Since our main objective was to minimize the false positive we find its performance satisfying.

Group 23 (Yuko Tsuchiya[1], Eiji Kanamori[2], Daron M. Standley[3], Haruki Nakamura[1], Kengo Kinoshita[4], [1]Institute for Protein Research, Osaka University, [2]Japan Biological Informatics Consortium, [3]Systems Immunology Lab, WPI Immunology Frontier Research Center (IFReC), Osaka University, [4]Graduate School of Information Sciences, Tohoku University)

Our scoring method is based on the interface complementarities in terms of the hydrophobicity and the electrostatic potential on the molecular surfaces of proteins and the shape of the surfaces. We use a linear combination of the degrees of complementarities for the three properties as a complementarity score of an interface,

$$SCR_{comp} = 0.343 \times H_{cmp} + 0.544 \times E_{cmp} + 0.112 \times S_{cmp} \qquad \text{Eq. 1}$$

where H_{cmp}, E_{cmp}, and S_{cmp} represent the degrees of complementarities for the hydrophobicity, the electrostatic potential and the shape, respectively. These complementarities are calculated as follows; (1) at each vertex on the molecular surface of each component protein of a complex, the hydrophobicity, the electrostatic potential, and the shape are calculated[62]. (2) In each interface, the numbers of complementary inter-subunit vertex pairs that have shorter distances than 3Å are counted for the three properties, respectively; N_{ele} is the number of complementary (positive and negative) vertex pairs for the electrostatic potential, N_{hyd} is that of hydrophobic and hydrophobic vertex pairs, N_{shape} is that of convex and concave vertex pairs, and N_{total} is the total number of all inter-subunit vertex pairs in an interface. (3) The ratio of the number of complementary vertex pairs to the total number of the vertex pairs in the interface, is defined as the complementarity for each property, such as $H_{cmp}=N_{hyd}/N_{total}$.

The optimization of the weights in Eq. 1 has been performed by using the docking models for the 74 representative hetero-dimers[63]. We prepared up to 500 models for each hetero-dimer by our docking method[64] as a training dataset, and optimized the three weights so that the discrimination between the native-like models that have a rmsd<10Å from the native complexes and the other models in the training set could be performed with maximum accuracy.

To add the effect of the size of the interface, we used the index that shows the fitness of interface surfaces, which is one of the terms of the score calculated in the docking[64]. The value of the cumulative distribution function for the index of the surface fitness is defined as an additional score, SCR_{surf}, which is calculated based on the values of the surface fitness of the native-like models in the training set.

We used the linear discriminant function of SCR_{comp} and SCR_{surf} as the final score, SCR_{final}. The function was constructed based on the SCR_{comp}s and the SCR_{surf}s of the native-like models and the other models in the training set.

$$SCR_{final} = 6.83(SCR_{comp} - 0.43) + 1.65(SCR_{surf} - 0.43) \qquad \text{Eq. 2}$$

In principle, this function judges the model with a positive score as a native or a native-like model, and that with a negative score as a non-native model. However, we empirically defined the five normalized categories as follows; the model with SCR_{final} =0.5 is categorized as “bind”, 0 = SCR_{final} <0.5 as “likely to bind”, -0.5= SCR_{final} <0 as “uncertain”, -1.0= SCR_{final} <-0.5 as “likely not to bind”, and SCR_{final} <-1.0 as “does not to bind”.

Group 24 (Camden M. Driggers[1], Robert G. Hall[2], Jessica L. Morgan[1] and Victor L. Hsu[1], [1]Department of Biochemistry and Biophysics, [2]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, OR, USA)

The designed protein complexes were binned based on a scoring function derived

from the protein-protein docking benchmark dataset[40]. The intermolecular and total energetic components of the 124 benchmark complexes and the 87 designed complexes were determined with HADDOCK[65]. These included terms corresponding to bond, angle, dihedral and improper energies, Lennard-Jones and electrostatic potentials, intermolecular van der Waals and electrostatic energies, desolvation energy and the buried surface area of the complex. Weighted values for these metric terms in the scoring function were empirically derived from a randomly selected subset of the benchmark dataset from which single clusters of acceptable complex structures were used (approximately 1000 structures including the selected crystal structures, the HADDOCK active and passive residues were determined by inspection of the respective crystal structures). These weighted values were refined by comparison to the corresponding values determined from a set of randomly generated (typically unacceptable) complex structures based on the same benchmark subset.

For the designed complexes, a short optimization was performed using HADDOCK, and each complex was binned into one of four binding categories based on its scoring: "binds" if score < 7.0; "likely to bind" if 7.0 = score < 8.0; "likely not to bind" if 8.0 = score < 9.0; "does not bind" if score = 9.0. Of the 87 designed complexes, 30 were determined to "bind", 18 as "likely to bind", 14 as "likely not to bind" and 25 as "does not bind". Of the 124 benchmark complexes, 90 were determined to "bind", 13 as "likely to bind", 13 as "likely not to bind" and 8 as "does not bind". Work is in progress to refine the weighted scoring function by incorporating polarity and structural and neighboring propensities using a machine learning approach, at which time the complex structures will be re-evaluated.

Group 26 (Jian Zhan, Yuedong Yang, and Yaoqi Zhou, Indiana University School of Informatics, Indiana University Purdue University at Indianapolus, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine)

Our main scoring function is based on a knowledge-based energy function with the distance-scaled finite ideal-gas reference state (DFIRE)[22] with a fine distance grid of 0.5Å[66]. Its application to predict binding affinity of protein-protein complexes is made by calculating the DFIRE energy for interacting atomic pairs at the interface[16]. The interface cutoff distance (7.0Å) was optimized by employing the dataset established

by Kastritis and Bonvin[32]. This dataset is a subset of Zdock Benchmark and contains 81 protein-protein complexes with known experimental binding affinities. The optimized correlation coefficient between predicted binding affinity and experimental binding affinity is 0.233. The DFIRE energy function was applied to 120 native complexes and 87 designed models relaxed by Rosetta and its Mathews correlation coefficient for separating native complexes from designed models is 0.485 with an optimized energy threshold (-4.0). To refine our prediction, we further employ the orientation components (OC) of the dipolar DFIRE (dDFIRE) potential function[67] that approximates polar atoms as points with directions defined by covalent bonds. This orientation component together with DFIRE separates targets into 4 regions that are scored as1 if EDFIRE<-4.0, EOC<-9.5, 2 if EDFIRE<-4.0, EOC>-9.5, 3 if EDFIRE>-4.0, EOC<-9.5 and 4 if EDFIRE>-4.0, EOC>-9.5. The threshold for the orientation component of dDFIRE was also from optimizing the Mathews correlation coefficient.

Group 28 (Panagiotis L. Kastritis and Alexandre M. J.J. Bonvin, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands)

For the purpose of this experiment, a physics-based potential was developed for predicting the binding affinity of the designed complexes and of those from the protein-protein docking benchmark. For all 207 complexes, a short optimization step was performed using the refinement interface of the HADDOCK web server[68] as previously described[32]. We then calculated theoretical dissociation constants ($-\log K_d$'s) for all complexes using the following equation:

$$-\log K_d = (w_a * E_{vdw} + w_b * E_{Elec} + w_c * E_{Hb} + w_d * ?G_{Desolv} + w_e * BSA) * F \quad (1)$$

where E_{vdw}, E_{Elec}, E_{Hb} denote the energetic contributions of the intermolecular van der Waals, the Electrostatics and the Hydrogen bonds. E_{vdw} denotes the Lennard-Jones potential, calculated with HADDOCK[65] as previously described[69]. E_{Elec} denotes the standard intermolecular Coulombic electrostatic potential with a distance-dependent dielectric term equal to $4r^5$, implemented in the FASTCONTACT algorithm[70]. E_{Hb} is a hydrogen bonding potential, originally developed in the Baker lab[24] and implemented in the FIREDOCK algorithm[71]. $?G_{Desolv}$ corresponds to the Lazaridis-Karplus solvation term[72],

calculated using ROSETTA[38]; it was calculated as the difference between sum of the solvation terms of the free chains and of the complex. BSA denotes the buried surface area in Å^2, calculated with HADDOCK2.1[69]. Finally, $w_a - w_e$ correspond to different weights for each contribution and F is a simple scaling factor (Kastritis & Bonvin, manuscript in preparation).

In order to assess the free energy difference between our theoretical calculations that derive from Equation (1) and the original experimental values, binding affinities referring to dissociation constants (K_d's, therefore units are in M) were converted into free energies of binding ($?_rG^0$), using the following equation:

$$?_rG^0=2.303RT*(-\log K_d) \quad (2)$$

The average error between the theoretical (predicted) and experimental binding free energies was expressed in *kcal/mol* as:

 (3)

where N represents the total number of complexes for which experimental binding affinities are available.

The multiple linear regression model in Equation (1) and its various weight factors was parameterized against a dataset of 81 experimentally determined protein-protein binding affinity data that we recently published[32]. During its development, we have found that it is not very sensitive to the optimization protocol of the structures for which the binding affinity is predicted: for example, complexes optimized using HADDOCK give very similar affinities to the ones optimized with Rosetta (R^2>0.80, average error = 1.3 *kcal/mol*). The Spearman correlation coefficient between the experimentally determined binding affinities and the theoretically calculated ones reach 0.54 (d.o.f. = 79) with an average error in their corresponding free energies of binding of 2.3 *kcal/mol* (on the 81 complexes from our benchmark). Note that since its publication, we discovered some errors in the reported binding affinities. If we

only include high-quality data from Isothermal Titration Calorimetry, Surface Plasmon Resonance and Spectrophotometric assays, the correlation significantly improves with an R=0.80 (d.o.f. = 45).

For the CAPRI round 21 experiment, we binned the 207 complexes in five categories according to their predicted binding affinities:
do not bind: $-\log K_d < 3$
likely not to bind: $3 = -\log K_d < 5$
uncertain: $5 = -\log K_d < 7$
likely to bind: $7 = -\log K_d < 9$
bind: $9 = -\log K_d$

Group 29 (Weiyi Zhang, Carlos J. Camacho, University of Pittsburgh, US)

We evaluated the co-crystals and models using *FastContact*[70; 73-76], one of the first free energy based scoring functions used to predict protein interactions. *FastContact* shows almost identical sensitivity and specificity rates when discriminating complex structures in the PDB regardless of whether one accounts for changes in van der Waals ($?E_{vdw}$) and/or internal ($?E_{int}$) energies, reflecting the optimal complementarity of protein-protein interactions. On the other hand, the simultaneous discrimination of both co-crystal and model structures improved by 20% with the addition of both $?E_{vdw}$ and $?E_{int}$, in the scoring function, reflecting the shortcomings of refining model structures.

In what follows, we refer to *FastContact* as the formula in Eq. 1 and *SmoothDock*[77] scoring as the formula in Eq. 2.

$$\nabla G = \nabla E^{elec} + \nabla G^{des} \qquad (1)$$

$$\Delta G = \Delta E_{elec} + \Delta G_{des} + \Delta E_{vdw} + \Delta E_{int} \qquad (2)$$

where
$?E_{elec}$ corresponds to the intermolecular Coulombic electrostatic potential with a distance-dependent constant equal to $4r$, and $?G_{des}$ is an empirical desolvation contact free energy that account for the hydrophobic interactions,

the self-energy change upon desolvating polar groups and the entropy loss of transferring a side chain from a protein surface to a bound conformation[41]. $?E_{vdw}$ and $?E_{int}$ correspond to the change in van der Vaals and internal energy upon binding, where *Evdw* and *Eint* are computed by CHARMM force field after 20X3 energy minimization using ABNR (adopted basis Newton-Raphson) steps and the CHARMM-19 potential with polar hydrogens only, distance-dependent dielectrics e = 4*r*, and fixed backbone. We note that our scoring function does not account for translational, rotational and vibrational entropies.

For the screening of ZDock2.0, ZDock3.0 and designed proteins, we set the binding free energy threshold for *FastContact* and *SmoothDock* at -21.6 kcal/mol and -79.0kcal/mol, respectively. Predicted sensitivity rates for ZDock datasets are 57.5% (69 true positives out of 120 total) and 58.33% (70 true positives out of 120) for *FastContact* and *SmoothDock* , respectively. On the other hand, specificity rates for designed models improve from a *FastContact* prediction of 68.97% (60 true negatives out of 87 total) to 88.51% (77 true negatives out of 87) when accounting for internal and solute van der Waals energies using *SmoothDock*. It is important to stress that our predictions do not involve any prior knowledge of protein-protein interactions, nor we made any attempt to incorporate features of the Rosetta scoring function in our analysis.

Group 30 (Krishna Praneeth Kilambi, Brian Weitzner, Justin Porter, Aroop Sircar and Jeffrey J. Gray)

We evaluated the following physical parameters for each protein from the set of designed and real complexes

1. Interface area per residue for each complex
2. Number of interface contacts per unit surface area of the complex
3. Solvent accessible surface area of the complex
4. Polar solvent accessible surface area of the complex
5. Van der Waals energy (the Rosetta attractive and the repulsive terms in the Lennard-Jones Potential were studied independently)
6. The solvation penalty (the Lazaridis-Karplus model as implemented in Rosetta)

The only parameter distribution that displayed noticeable distinction between the natural and the designed complexes was the interface area per residue for each of the complexes. On an average, the designed complexes were found to have smaller Iarea/nres values. The distributions of the all the other parameters showed no significant separation between the native and designed complexes.

Group 31 (Masahito Ohue, Nobuyuki Uchikoga, Yuri Matsuzaki, Takashi Ishida and Yutaka Akiyama, Tokyo Institute of Technology, Japan)

We predicted PPIs by a method that utilized rigid body protein-protein docking[78]. We divided each complex structure into 2 protein chains and re-docked them by using our docking system named MEGADOCK. Finally, we judged whether they interact or not based on the normalized docking score.

MEGADOCK is based on the FFT calculation like ZDOCK[79]. MEGADOCK has the original shape complementarity scoring model called real Pairwise Shape Complementarity(rPSC) and CHARMm19 electrostatic model. ZDOCK Benchmark complexes are re-docked by MEGADOCK without changing them. Designs complexes are divided into chain A and B, then re-docked by MEGADOCK. We generated the 3600x3=10800 decoys that include top 3 decoys for each of ligand orientations (3600 with 15 degree intervals).

Then, we applied ZRANK[36]m for the decoys to exclude physicochemically unrealistic models and selected the best 2000 models. The re-ranked decoys' MEGADOCK scores are converted to *Z*-score, as follows:

$Z_i = \frac{S_i - m}{s}$ *Si* is *i*-th decoy's MEGADOCK score, μ is the mean of all decoys' MEGADOCK scores and *s* is the standard deviation of the scores. Then we calculated the "Raw score" *E* as follows: $E = \begin{cases} Z_k & \text{if } Z_k > 6 \text{ and } Z_k \text{ is maximum value} \\ Z_1 & \text{otherwise} \end{cases}$

Finally, the Raw score is normalized, using the following scheme; 1. binds (*E*

> 4); 2. likely to bind (4 >= *E* > 2.5); 3. likely not to bind (2.5 >= *E* > 1.5); 4. Don't bind (1.5 >= *E*). 5. uncertain is not treated in this method.

Group 32 (Raed Khashan, Stephen Bush, Denis Fouches, and Alexander Tropsha, University of North Carolina at Chapel Hill)

Our method is centered on a simple knowledge-based scoring function that utilizes frequent geometric patterns of interacting residues found at the interfaces of X-ray characterized protein-protein complexes.

The approach includes the following steps. First, protein-protein interfaces of each complex in the X-ray crystallographic native complexes (we used Vakser and co-workers' Bound Dockground[80]) are represented by labeled graphs where nodes are residues' centroids and edges connect centroids located within certain distance (we used 10 A°) of each other. These interfacial residues were identified using Almost Delaunay tessellation[81], therefore allowing some flexibility in the selection process due to variations in residues positions in low resolution crystallographic complexes. Second, efficient subgraph mining techniques are used to find frequent subgraphs that occur in no less than a certain percentage of the native complexes; these frequent subgraphs (or patterns of interacting residues) identify structural motifs that we regard as "classical" interacting patterns.

Thus, given a test set of protein-protein complexes, they can be scored based on the interaction patterns found at their interface that match these "classical" frequent patterns. The scoring function takes into account the frequency of the matching "classical" patterns in the native complexes, their size, and the degree of geometrical similarity between patterns in the test proteins and their matching "classical" patterns. The scoring function takes also into account the number and ratio of interacting residues at test proteins interface that found a match with the "classical" patterns. These factors can be used to derive the following formula for the scoring function:

$$\text{Score} = \| \sum_{N} \sum_{M} |Pi|/RMSDijpattern\| + \|X1\| + \|X2\| + \|X3\| + \|X4\|$$

i j

Where N is the total number of frequent (”classical”) patterns found at the interface, M is the frequency of the pattern i in the set of native complexes, and therefore represents the number of modes of interaction (number of different internal geometric coordinate sets) for that pattern. |Pi| is the size of the pattern Pi (i.e., total number of protein residues in the pattern), and RMSDijpattern is calculated for the best fit between pattern Pi in the test complex and the matching “classical” pattern. The first summation is over all patterns that are found at the interface. The second summation reflects the frequency of each pattern and the different modes of interaction for each pattern. Also, to avoid dividing by zero, an epsilon value of 1*10-60 is added to the RMSDpattern. (This value is chosen based on the smallest empirical RMSDpattern value that was found in our studies.) Other parameters used: X1 is the number of interfacial residues found a match with the classical patterns. X2 is the ratio of interfacial residues that found a match with the classical patterns. X3 is the number of classical patterns found at the interface. Finally, X4 is the number of classical patterns found at the interface divided by the number of interfacial residues found a match; i.e., the average number of patterns matched per one interfacial residue. Therefore, based on the formula, one can conclude that the higher the score, the closer the test complex to its native structure.

Group 33 (Juan Esquivel-Rodriguez, Daisuke Kihara, Purdue University)
The scoring function we used is a linear combination of nine physics-based and a knowledge-based potential. The first two terms are the repulsive and the attractive parts of the van der Waals potential using the 12-6 Lennard Jones potential. The next four terms are the electrostatic potential, which are split into repulsive and attractive as well as long and short range terms. The other three terms are a hydrogen bonding term, a solvation energy by Lazaridis and Karplus[72] and a knowledge-based atom contact potential[41]. Weighting factors of the terms were trained on the decoy set of the ZDOCK benchmark 2.0[79] and a separate set of decoys generated by running ZDOCK on the dataset used by Huang & Zou[82]. Genetic algorithms and the linear

regression were used separately, and the best performing model overall was selected to generate the final scoring function.

Group 35 (PB Stranges, R Jacak and B Kuhlman. University of North Carolina Chapel Hill)

To discriminate between the native protein-protein interfaces and the designed ones, we chose to use a set of four metrics that can be computed from the structures: predicted binding energy per interface area, number of unsatisfied hydrogen bonds per interface area, ratio of hydrogen bond energy to total binding energy and the RosettaHoles score[83]. All designed and native structures used in this analysis were repacked and minimized using the Rosetta energy function[84].

The predicted binding energy was calculated by taking the difference in energy between the structure of the complex and the separated binding partners. The area of the interface was computed in the same manner. The first metric, dG/dSASA, was obtained by dividing the binding energy by the area of the interface. The second metric, Unsat/dSASA, is the number of buried-unsatisfied polar atoms located at the interface divided by the area of the interface. The third metric, HBond energy/dG, is the proportion of hydrogen bond energy to the total binding energy.

A normal distribution was fit to each of the described metrics for native structures with resolution better than or equal to 2.2 angstroms. The corresponding cumulative distribution function for each metric was used to represent a score for how well a structure compares to natives. For each metric a score of 1.0 represents an above average comparison to natives while a score of 0.0 represents no correspondence to natives.

The final metric is the RosettaHoles score. The RosettaHoles score represents the probability that a set of solvent inaccessible voids comes from a high-resolution crystal structure. Voids are determined by finding the largest spherical hole adjacent to all buried atoms, pruning away solvent accessible regions, and then clustering the holes into contiguous cavities. The scores for this metric are similar to the previous three, with 1.0 being ideal and 0.0 being completely unlike native proteins.

The scores for each of the four metrics for all native and designed structures were summed to yield the final raw score. This score ranged between 0.0 and 4.0 with 0.0 being no correspondence to native interfaces (does not bind) and 4.0 being as good as or better than native interfaces (binds). The normalized scores are as follows where RS represents the raw score: 4 (does not bind) for RS <1.5; 3 (likely not to bind) for 1.5<RS<2.0; 2 (likely to bind) for 2.0<RS<2.5; 1 (binds) for RS>2.5.

Some native structures were not evaluated due to problems obtaining a suitable minimized structure.

Group 36 (Sheng -You Huang, Xiaoqin Zou, University of Missouri-Columbia)

ITScore/PP[82] was used to evaluate the 87 designed complex models from the Baker lab and 120 native complexes of the CAPRI Target T45. There was no parameter optimization or training in the pair interaction potentials of ITScore/PP for these protein-protein complexes.
ITScore/PP is an all-atomic distance-dependent knowledge-based scoring function derived from a physics-based iterative method that circumvents the long-standing reference state problem in the knowledge-based/statistical approaches. The basic idea of the method is to improve a set of effective pair potentials by iteration until the derived potentials can reproduce the atomic pair distribution functions of the experimentally determined complex structures in a diverse training set that is different from the current test sets of complexes [85; 86]. A second advantage of ITScore/PP is that the derivation considers the whole binding energy landscapes of the complexes by including both the native structures and decoys according to a Boltzmann probability, rather than considering only the energy minima (i.e., native structures) as done in conventional knowledge-based scoring functions[87]. The pair interaction potentials in ITScore/PP were derived based on 20 heavy atom types and 851 non-redundant, biological protein-protein complex structures[82].

During the binding score calculations for each protein-protein complex in the benchmarks, both protein partners were treated as rigid bodies and were allowed for local minimization in their coordinates.

Supplemental Tables

PDB ID	Buried surface area	Computed binding energy
Native complexes		
1A2K	1640	-19.1
1ACB	1650	-28.3
1AHW	2060	-24.2
1AK4	1000	-17.8
1AKJ	1990	-17.7
1AVX	1730	-14.7
1AY7	1310	-16.8
1AZS	1980	-23.7
1B6C	1920	-15.7
1BGX	6440	-1.6
1BJ1	1810	-28.2
1BKD	3250	-39.4
1BUH	1390	-14.5
1BVK	1250	-13.0
1BVN	2250	-19.2
1CGI	2080	-21.8
1D6R	1410	-17.6
1DE4	2340	-5.7
1DFJ	2540	-23.0
1DQJ	1780	-16.1
1E4K	1580	5.4
1E6J	1250	-16.7
1E96	1240	-16.7
1EAW	1940	-10.3
1EER	2240	-29.3
1EFN	1310	-10.2
1EZU	2710	-40.6
1F34	3350	-32.1
1F51	2520	-13.3
1FAK	3150	-32.0
1FC2	1240	-7.4
1FQ1	1840	-13.1
1FQJ	1960	-20.8
1FSK	1730	-23.0
1GCQ	1280	-18.1
1GHQ	750	-9.2
1GLA	1410	-13.8
1GP2	2310	-22.8
1GPW	2260	-20.3
1GRN	2330	-23.5
1H1V	2080	-20.1
1HE1	2080	-19.0
1HE8	1480	-13.9
1HIA	1800	-24.2

1I2M	2940	-33.4
1I4D	1710	-12.4
1I9R	1600	-9.6
1IB1	2930	-30.7
1IBR	3640	-24.1
1IJK	1760	-6.4
1IQD	2110	-30.2
1IRA	3500	-27.3
1J2J	1200	-15.9
1JMO	3880	-39.7
1JPS	1990	-25.3
1K4C	1540	-28.0
1K5D	2820	-21.3
1K74	1260	-9.7
1KAC	1540	-12.2
1KKL	1650	-10.2
1KLU	1260	-13.1
1KTZ	1000	-15.8
1KXP	3600	-32.4
1KXQ	2170	-28.5
1M10	2210	-23.4
1MAH	2070	-26.1
1ML0	2440	-33.0
1MLC	1460	-15.4
1N2C	3990	-24.4
1N8O	1920	-31.4
1NCA	2030	-17.9
1NSN	1860	-12.7
1NW9	2050	-32.3
1OPH	1390	-18.6
1PPE	1750	-27.1
1PXV	2400	-39.0
1QA9	1260	-13.0
1QFW	1580	-11.8
1R0R	1440	-29.8
1R8S	2880	-39.1
1RLB	1580	-6.9
1S1Q	1340	-15.0
1SBB	1250	-16.4
1T6B	1910	-15.2
1UDI	2070	-25.3
1VFB	1400	-20.2
1WEJ	1200	-15.9
1WQ1	3120	-15.7
1XD3	2280	-33.6
1XQS	2400	-23.5
1Y64	2510	-11.3
1YVB	1730	-16.4
1Z0K	1880	-23.9
1Z5Y	1370	-22.7

1ZHI	1300	-18.3
2AJF	1800	-15.4
2B42	2680	-23.9
2C0L	1930	-19.4
2CFH	2320	-31.3
2FD6	1150	-19.1
2H7V	1570	-9.9
2HLE	2190	-28.6
2HMI	1310	-7.2
2HQS	2410	-28.4
2HRK	1560	-13.8
2I25	1450	-26.1
2JEL	1620	-15.0
2MTA	1480	-12.4
2NZ8	2590	-29.1
2O8V	1680	See note 1
2OOB	800	-13.2
2OT3	2380	-40.6
2PCC	1190	-5.8
2QFW	1580	-11.8
2SIC	1710	-23.3
2SNI	1660	-25.9
2UUY	1380	-16.9
2VIS	1450	-13.7
7CEI	1440	-11.5
Designed complexes		
design_1	1200	-16.4
design_2	1200	-13.9
design_3	1640	-23.4
design_4	1250	-17.7
design_5	1240	-14.2
design_6	1230	-16.0
design_7	1490	-21.0
design_8	1280	-15.1
design_9	1110	-17.9
design_10	1330	-16.5
design_11	1360	-19.4
design_12	1730	-18.7
design_13	1320	-20.4
design_14	1110	-16.5
design_15	1210	-17.3
design_16	1330	-16.8
design_17	1230	-17.9
design_18	1330	-17.7
design_19	1390	-15.4
design_20	1930	-23.8
design_21	1230	-16.7
design_22	1130	-18.4
design_23	1300	-20.9
design_24	1300	-12.6

design_25	1030	-12.9
design_26	1300	-21.9
design_27	1080	-19.6
design_28	1080	-14.6
design_29	1260	-21.5
design_30	1630	-21.4
design_31	1040	-14.0
design_32	870	-14.9
design_33	1480	-16.6
design_34	1730	-20.8
design_35	1760	-24.8
design_36	1380	-15.0
design_37	1500	-17.5
design_38	1520	-24.4
design_39	1280	-14.0
design_40	1230	-15.9
design_41	1260	-15.9
design_42	1410	-17.8
design_43	1340	-20.5
design_44	1510	-20.5
design_45	1580	-18.2
design_46	1070	-15.2
design_47	1320	-19.1
design_48	1480	-18.7
design_49	1150	-16.5
design_50	1160	-16.9
design_51	1240	-15.1
design_52	1750	-16.4
design_53	1500	-21.7
design_54	1180	-17.1
design_55	1180	-12.3
design_56	1130	-17.1
design_57	1430	-21.4
design_58	1460	-17.8
design_59	1350	-22.8
design_60	1120	-13.4
design_61	1650	-13.2
design_62	1440	-17.8
design_63	1290	-14.4
design_64	1100	-11.3
design_65	1510	-22.4
design_66	1420	-17.0
design_67	1230	-12.6
design_68	1200	-13.9
design_69	1520	-19.8
design_70	1180	-18.1
design_71	1410	-11.3
design_72	1780	-21.9
design_73	950	-15.6
design_74	980	-16.2

design_75	1170	-13.8
design_76	1590	-19.0
design_77	1640	-16.4
design_78	1080	-15.5
design_79	1070	-15.5
design_80	1060	-16.5
design_81	1490	-18.9
design_82	1400	-15.5
design_83	1690	-22.5
design_84	1420	-17.6
design_85	1450	-20.7
design_86	1170	-12.4
design_87	1180	-20.0

Table S1: Buried surface area ($Å^2$) and computed binding affinity (Rosetta energy units; R.e.u.) for the set of native and designed complexes. Values for surface area and binding energy were rounded to 10 $Å^2$ and 0.1R.e.u., respectively.

1 Complex of a covalently linked interface, confounding binding energy calculations.

Table S2

1FC2 1KAC 1IJK 2FD6 1EFN 1E6J 1Z5Y 1KLU 2VIS 1ZHI 1S1Q
1ACB 1E4K 1RLB 1GCQ 2PCC 2HMI 1SBB 1AK4 2MTA 1KTZ 1GLA
1GHQ 1J2J 2OOB

25 PDB entries for hydrophobic interfaces from the docking benchmark. This list of structures has the lowest computed desolvation penalty upon binding in the docking benchmark.

Table S3

Group	AUC
1	64
2	83
4	64
5	69
6	74
7	81
8	64
9	61
10	79
11	53
12	72
14	53
16	59
17	63
20	75
21	62
22	51
23	75
24	73
26	56
28	55
29	51
30	56
31	67
32	68
33	57
35	74
36	60

Area under the curve (AUC) percentages of participants in discriminating designed complexes from a subset of hydrophobic natural complexes (Table S2). The performance of most groups is worse against this subset of hydrophobic complexes than against the entire docking benchmark (Fig. 2).

Supplemental Figures

Figure S1 Per-group classification of designed and natural interfaces according to their binding propensity.

Figure S2: Misclassification of an active design as a non-binder. Design 45 was experimentally tested and shown to bind its target after the benchmark was completed by the participants[2], providing a blind test of the metrics. None of the groups predicted that this complex would bind.

Figure S1

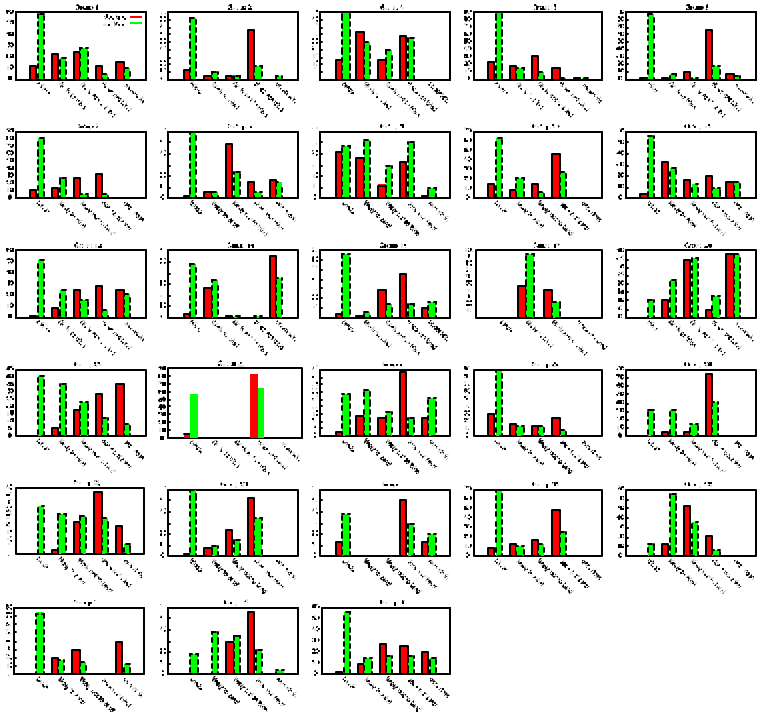
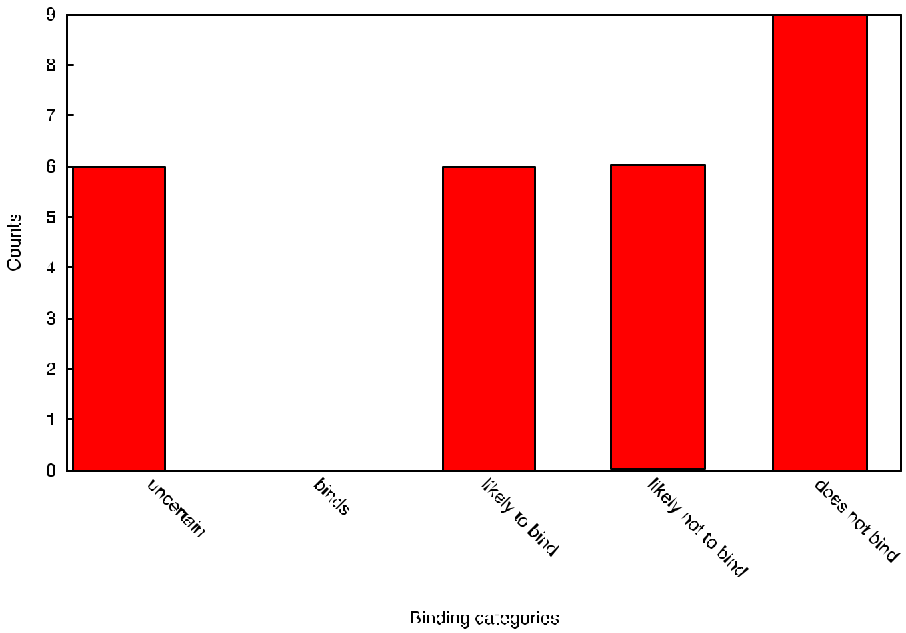


Figure S2


Counts
0
1
2
3
4
5
6
7
8
9
uncertain
binds
likely to bind
likely not to bind
does not bind
Binding categories

Supplemental References

1. 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**, 755-68.
2. Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E.-M., Wilson, I. A. & Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816-821.
3. Studier, F. W. (2005). Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* **41**, 207-34.
4. Park, H. S., Nam, S. H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J. & Kim, H. S. (2006). Design and evolution of new catalytic activity with an existing protein scaffold. *Science* **311**, 535-538.
5. Alsallaq, R. & Zhou, H. X. (2008). Electrostatic rate enhancement and transient complex of protein-protein association. *Proteins* **71**, 320-35.
6. Schreiber, G., Haran, G. & Zhou, H. X. (2009). Fundamental aspects of protein-protein association kinetics. *Chem Rev* **109**, 839-60.
7. Qin, S. & Zhou, H. X. (2009). Dissection of the high rate constant for the binding of a ribotoxin to the ribosome. *Proc Natl Acad Sci U S A* **106**, 6974-9.
8. Demerdash, O. N., Buyan, A. & Mitchell, J. C. (2010). ReplicOpter: a replicate optimizer for flexible docking. *Proteins* **78**, 3156-65.
9. Terashi, G., Takeda-Shitaka, M., Kanou, K., Iwadate, M., Takaya, D., Hosoi, A., Ohta, K. & Umeyama, H. (2007). Fams-ace: a combined method to select the best model after remodeling all server models. *Proteins* **69 Suppl 8**, 98-107.
10. Terashi, G., Takeda-Shitaka, M., Kanou, K., Iwadate, M., Takaya, D. & Umeyama, H. (2007). The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation. *Proteins* **69**, 866-72.
11. Chen, R., Li, L. & Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80-7.
12. Moal, I. H. & Bates, P. A. (2010). SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Int J Mol Sci* **11**, 3623-48.
13. Chaudhury, S., Lyskov, S. & Gray, J. J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689-91.
14. Li, X., Moal, I. H. & Bates, P. A. (2010). Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding. *Proteins* **78**, 3189-96.
15. Shaefer, M. & Karplus, M. (1996). A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Phys. Chem.* **100**, 1578-1599.
16. Liu, S., Zhang, C., Zhou, H. & Zhou, Y. (2004). A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins* **56**, 93-101.
17. Mitra, P. & Pal, D. (2010). New measures for estimating surface complementarity and packing at protein-protein interfaces. *FEBS Lett* **584**, 1163-8.

18. Im, W., Lee, M. S. & Brooks, C. L., 3rd. (2003). Generalized born model with a simple smoothing function. *J Comput Chem* **24**, 1691-702.
19. 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**, 482-91.
20. Zacharias, M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci* **12**, 1271-82.
21. Fiorucci, S. & Zacharias, M. (2010). Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins* **78**, 3131-9.
22. Zhou, H. & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* **11**, 2714-26.
23. Zhou, H. & Zhou, Y. (2002). Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. *Proteins* **49**, 483-92.
24. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**, 1239-1259.
25. Liang, S., Meroueh, S. O., Wang, G., Qiu, C. & Zhou, Y. (2009). Consensus scoring for enriching near-native structures from protein-protein docking decoys. *Proteins* **75**, 397-403.
26. Cole, C. & Warwicker, J. (2002). Side-chain conformational entropy at protein-protein interfaces. *Protein Sci* **11**, 2860-70.
27. Su, Y., Zhou, A., Xia, X., Li, W. & Sun, Z. (2009). Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. *Protein Sci* **18**, 2550-8.
28. Bourquard, T., Bernauer, J., Azé, J. & A., P. (2009). *Sixth International Symposium on Voronoi Diagrams.*
29. Bernauer, J., Aze, J., Janin, J. & Poupon, A. (2007). A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics* **23**, 555-62.
30. Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J. & Poupon, A. (2008). DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics* **24**, 652-8.
31. Bernauer, J., Poupon, A., Aze, J. & Janin, J. (2005). A docking analysis of the statistical physics of protein-protein recognition. *Phys Biol* **2**, S17-23.
32. Kastritis, P. L. & Bonvin, A. M. (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J Proteome Res* **9**, 2216-25.
33. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O. & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* **80**, 505-15.
34. Eyal, E., Yang, L. W. & Bahar, I. (2006). Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* **22**, 2619-27.
35. Bruschweiler, R. (1995). Collective protein dynamics and nuclear spin relaxation. *J. Chem. Phys.* **102**, 3396-3403.
36. Pierce, B. & Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins* **67**, 1078-86.

37. Pierce, B. & Weng, Z. (2008). A combination of rescoring and refinement significantly improves protein docking performance. *Proteins* **72**, 270-9.
38. 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**, 281-99.
39. Cheng, T. M., Blundell, T. L. & Fernandez-Recio, J. (2007). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins* **68**, 503-15.
40. Hwang, H., Pierce, B., Mintseris, J., Janin, J. & Weng, Z. (2008). Protein-protein docking benchmark version 3.0. *Proteins* **73**, 705-9.
41. Zhang, C., Vasmatzis, G., Cornette, J. L. & DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol* **267**, 707-26.
42. Gong, X., Wang, P., Yang, F., Chang, S., Liu, B., He, H., Cao, L., Xu, X., Li, C., Chen, W. & Wang, C. (2010). Protein-protein docking with binding site patch prediction and network-based terms enhanced combinatorial scoring. *Proteins* **78**, 3150-5.
43. Chang, S., Gong, X., Jiao, X., Li, C., Chen, W. & Wang, C. (2010). Network analysis of protein-protein interaction. *Chinese Science Bulletin* **9**, 814-822.
44. Moont, G., Gabb, H. A. & Sternberg, M. J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35**, 364-73.
45. Guharoy, M. & Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci U S A* **102**, 15447-52.
46. Schneider, R., de Daruvar, A. & Sander, C. (1997). The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res* **25**, 226-30.
47. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-10.
48. Chakrabarti, P. & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins* **47**, 334-43.
49. Guharoy, M. & Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. USA* **102**, 15447-15452.
50. Liu, S. & Vakser, I. A. (2011). DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC Bioinformatics* **12**, 280.
51. Samudrala, R. & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* **275**, 895-916.
52. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* **65**, 392-406.
53. Chuang, G. Y., Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. (2008). DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophys J* **95**, 4217-27.
54. Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J* **85**, 1145-64.
55. Tovchigrechko, A. & Vakser, I. A. (2005). Development and testing of an automated approach to protein docking. *Proteins* **60**, 296-301.

56. London, N. & Schueler-Furman, O. (2008). Funnel hunting in a rough terrain: learning and discriminating native energy funnels. *Structure* **16**, 269-79.
57. Le Grand, S. M. & Merz, K. M. (1993). Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. *J Comput Chem* **14**, 349-352.
58. Zhu, H., Domingues, F. S., Sommer, I. & Lengauer, T. (2006). NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* **7**, 27.
59. Wang, C., Schueler-Furman, O., Andre, I., London, N., Fleishman, S. J., Bradley, P., Qian, B. & Baker, D. (2007). RosettaDock in CAPRI rounds 6-12. *Proteins* **69**, 758-63.
60. Schreiber, G., Shaul, Y. & Gottschalk, K. E. (2006). Electrostatic design of protein-protein association rates. *Methods Mol Biol* **340**, 235-49.
61. Potapov, V., Cohen, M., Inbar, Y. & Schreiber, G. (2010). Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. *BMC Bioinformatics* **11**, 374.
62. Tsuchiya, Y., Kinoshita, K. & Nakamura, H. (2006). Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity. *Protein Eng Des Sel* **19**, 421-9.
63. Tsuchiya, Y., Kanamori, E., Nakamura, H. & Kinoshita, K. (2009). Classification of heterodimer interfaces using docking models and construction of scoring functions for the complex structure prediction. *Advances and Applications in Bioinformatics and Chemistry* **2**, 79-100.
64. Kanamori, E., Murakami, Y., Tsuchiya, Y., Standley, D. M., Nakamura, H. & Kinoshita, K. (2007). Docking of protein molecular surfaces with evolutionary trace analysis. *Proteins* **69**, 832-8.
65. Dominguez, C., Boelens, R. & Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731-7.
66. Yang, Y. & Zhou, Y. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci* **17**, 1212-9.
67. Yang, Y. & Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* **72**, 793-803.
68. de Vries, S. J., van Dijk, M. & Bonvin, A. M. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* **5**, 883-97.
69. de Vries, S. J., van Dijk, A. D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T. & Bonvin, A. M. (2007). HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **69**, 726-33.
70. Camacho, C. J. & Zhang, C. (2005). FastContact: rapid estimate of contact and binding free energies. *Bioinformatics* **21**, 2534-6.
71. Andrusier, N., Nussinov, R. & Wolfson, H. J. (2007). FireDock: fast interaction refinement in molecular docking. *Proteins* **69**, 139-59.
72. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133-52.
73. Camacho, C. J. & Gatchell, D. W. (2003). Successful discrimination of protein interactions. *Proteins* **52**, 92-7.

74. Camacho, C. J., Gatchell, D. W., Kimura, S. R. & Vajda, S. (2000). Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins* **40**, 525-37.
75. Camacho, C. J., Ma, H. & Champ, P. C. (2006). Scoring a diverse set of high-quality docked conformations: a metascore based on electrostatic and desolvation interactions. *Proteins* **63**, 868-77.
76. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. (2004). ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res* **32**, W96-9.
77. Camacho, C. J. & Vajda, S. (2001). Protein docking along smooth association pathways. *Proc Natl Acad Sci U S A* **98**, 10636-41.
78. Matsuzaki, Y., Matsuzaki, Y., Sato, T. & Akiyama, Y. (2009). In silico screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis. *J Bioinform Comput Biol* **7**, 991-1012.
79. Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R. & Weng, Z. (2007). Integrating statistical pair potentials into protein complex prediction. *Proteins* **69**, 511-20.
80. Douguet, D., Chen, H. C., Tovchigrechko, A. & Vakser, I. A. (2006). DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics* **22**, 2612-8.
81. Bandyopadhyay, D. & Snoeyink, J. (2004). Almost-Delaunay simplices: nearest neighbor relations for imprecise point. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms,*, 403-412.
82. Huang, S. Y. & Zou, X. (2008). An iterative knowledge-based scoring function for protein-protein recognition. *Proteins* **72**, 557-79.
83. Sheffler, W. & Baker, D. (2009). RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci* **18**, 229-39.
84. Das, R. & Baker, D. (2008). Macromolecular modeling with rosetta. *Annu Rev Biochem* **77**, 363-82.
85. Huang, S.-Y. & Zou, X. (2006). An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J Comput Chem* **27**.
86. Huang, S.-Y. & Zou, X. (2006). An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J Comput Chem* **27**, 1865-1875.
87. Huang, S.-Y. & Zou, X. (2010). Mean-force scoring functions for protein-ligand binding. *Annu Rep Comput Chem* **6**.