# Journal of Proteome research

# Are Scoring Functions in Protein−Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark

**Panagiotis L. Kastritis and Alexandre M. J. J. Bonvin***

*Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, 3584CH, Utrecht, The Netherlands*

The design of an ideal scoring function for protein−protein docking that would also predict the binding affinity of a complex is one of the challenges in structural proteomics. Such a scoring function would open the route to *in silico*, large-scale annotation and prediction of complete interactomes. Here we present a protein−protein binding affinity benchmark consisting of binding constants ($K_d$'s) for 81 complexes. This benchmark was used to assess the performance of nine commonly used scoring algorithms along with a free-energy prediction algorithm in their ability to predicting binding affinities. Our results reveal a poor correlation between binding affinity and scores for all algorithms tested. However, the diversity and validity of the benchmark is highlighted when binding affinity data are categorized according to the methodology by which they were determined. By further classifying the complexes into low, medium and high affinity groups, significant correlations emerge, some of which are retained after dividing the data into more classes, showing the robustness of these correlations. Despite this, accurate prediction of binding affinity remains outside our reach due to the large associated standard deviations of the average score within each group. All the above-mentioned observations indicate that improvements of existing scoring functions or design of new consensus tools will be required for accurate prediction of the binding affinity of a given protein−protein complex. The benchmark developed in this work will serve as an indispensable source to reach this goal.

**Keywords:** Biomolecular complexes • interactome • energetics • correlation • proteomics

## Introduction

Understanding protein−protein interactions is crucial step in the investigation of intracellular signaling pathways,[1] antibody−antigen pairing,[2] and enzyme−inhibitor interactions.[3] Although current structural biology tools have broaden our knowledge in single protein structure, function and dynamics,[4] the situation differs substantially in the case of protein−protein complexes: Due to experimental limitations in probing protein−protein interactions[5] and solving the structure of biomolecular complexes,[6] complementary computational approaches are often needed to assist experimentalists in investigating how two proteins of known structure interact and form a three-dimensional (3D) complex. Protein−protein docking algorithms have been developed for this purpose. They use geometric, steric and energetic considerations to predict the atomic structure of a complex.[7] Every docking program incorporates two key parts:[8] The search algorithm that samples configurational and conformational degrees of freedom and the scoring function that ranks the solutions. Even if docking methods are highly promising tools for the modeling of multiprotein complexes,[9,10] they do not (yet) allow a reliable estimation of the binding affinity of the complex.[11] For protein−protein complexes, affinity can be described by dissociation equilibrium constants ($K_d$'s) that span 10 orders of magnitude ($10^{-14} > K_d > 10^{-5}$). These values correspond to Gibbs free-energy changes upon binding ($\Delta G^\theta = -RT \ln K_d$) in between 7 and 20 kcal/mol.

What if these algorithms were able to predict the binding affinity of a macromolecular complex? Such a development would provide a quantitative description of the actual interactions among complexes. Interactomes obtained from high-throughput experimental methods could be cleaned up by removing false positive pairs based on computational binding affinity predictions, and even *in silico* interactome prediction might come within reach. Furthermore, coupling the prediction of structure, dynamics and binding affinity of protein−protein complexes would permit a time-scale simulation of the system, highlighting systemic/network features.[12]

Over the years, much effort has been placed in predicting the binding affinity of biomolecular complexes: Sophisticated approaches for estimating free-energy contributions have been reported[13−17] including Monte Carlo Conformational searches,[13] free-energy perturbations,[14] Poisson−Boltzmann,[15] generalized-Born solvation,[16] and atomic continuum electrostatic calculations.[17] All these methods are however computationally demanding and cannot be used for free-energy screening and binding affinity prediction in the context of

* To whom correspondence should be addressed. Prof. Dr. Alexandre M. J. J. Bonvin, NMR Spectroscopy Group, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands, E-mail: a.m.j.j.bonvin@uu.nl. Telephone: +310302533859. Fax: +31030253762.

protein−protein docking where thousands of models might need to be evaluated. Alternative, simplified methods that mostly relate to changes in the solvent accessible surface area upon binding have been proposed instead.[18−21] Some of these proved not to be ubiquitously applicable,[22,23] whereas most of them have not yet been thoroughly validated.[12,21] Some promising new methods have been published[21] that show very good results despite a limited training set (up to 24 protein−protein and protein−peptide complexes with known $K_d$ values).

Scoring functions used in docking have been also proposed for binding affinity prediction.[24−27] However, again, these have not yet been tested against a large number of complexes with known binding affinity, basically due to the absence of such a data set. In 2008, the protein−protein docking benchmark 3.0 was published.[28] It includes experimentally determined structures in both their bound and unbound form for 124 complexes. This benchmark is widely used for development and testing of docking algorithms.[29−35] It does not, however, include any binding affinity data. A benchmark including such values would be an indispensable tool for evaluating current protein−protein scoring functions, and would offer a common ground for the development and optimization of algorithms for the prediction of binding affinities.

We present here a novel Binding Affinity Benchmark that includes $K_d$ constants of 81 protein complexes together with the method used to determine them. The performance of nine commonly used scoring functions and one very fast free-energy prediction algorithm for protein−protein complexes are tested against this benchmark for their ability (or lack of) to predict binding affinities of protein−protein complexes.

## Materials and Methods

**Data Set.** The Protein−Protein Docking Benchmark 3.0[28] was used as a data set for this work. Literature searches were performed in order to find the experimentally determined $K_d$'s of the complexes. $K_d$'s could be found for 81 out of 124 complexes. Other biochemical data that relate to the binding affinity were also collected such as pH, temperature, and the experimental method used to measure the $K_d$'s.

**Optimization of the Experimentally Determined 3D-Structures of the Protein−Protein Complexes.** All 81 complexes (bound form) were optimized using the HADDOCK Web server (http://haddock.chem.uu.nl). HADDOCK is a high-performing docking approach[36] that uses a search algorithm in which experimental data can be incorporated and drive the docking procedure.[37] The server was however not used for docking of the bound forms of the molecules, but only to refine the given complex and consequently calculate the HADDOCK score. The complexes were optimized either by energy minimization or by a gentle molecular dynamics refinement, both in explicit solvent. Detailed methodology can be found in the Supporting Information. To check for a potential bias introduced by the energy minimization within HADDOCK, scoring of the original crystal structures was also performed without further optimization.

**Scoring.** The scores and their components were calculated for each complex (original, energy minimized and water-refined) with ATTRACT,[38] DFIRE,[39] FASTCONTACT,[40] FIREDOCK,[41] HADDOCK,[36] PISA,[42] PYDOCK,[43] ROSETTA-DOCK[27] and ZRANK.[44]

Affinity prediction was performed with AFFINITY-SCORE1.0.[21] The free-energy predictor was kindly provided by CMD Bioscience (http://www.cmdbioscience.com). The pro-gram uses a seven-term linear regression model, calculating total number of solvent exposed charged (a) and hydrophobic (b) atoms, number of hydrogen bonds (c), net number of short-range ($\leq 4$ Å) charge−charge or salt bridge interactions across the protein−protein interface (d), interface gap or void volume (e), change in the number of solvent exposed side-chain torsions or the total number of side chain torsions buried at the interface (f) and a constant contribution (g), giving them a corresponding weight.

The HADDOCK score is a consensus term that includes van der Waals and Electrostatic interactions and a Desolvation term. These values are weighted and combined into the HADDOCK score. The latter was obtained from the HADDOCK Web server (see above).

PYDOCK also uses a weighted score, including energy terms similar to HADDOCK (van der Waals, electrostatics interactions and desolvation energy); the weighting of the terms is, however, different.

The ATTRACT score is a consensus term of van der Waals interactions and Electrostatics calculated using a coarse-grained model of the protein−protein complex.

PISA analyses and predicts structural and chemical properties of macromolecular interfaces, as well as their probable dissociation pattern. Theoretical values that describe energy terms (for example, free energy of dissociation) are also calculated. The scores were obtained from the PISA server at http://www.ebi.ac.uk/msd-srv/prot_int/.

FIREDOCK rescores rigid-body protein−protein docking solutions. The refinement is restricted to interface side-chain rearrangement and to soft rigid-body optimization. The score includes Atomic Contact Energy, softened van der Waals interactions, partial electrostatics and additional estimations of the binding free energy. They were obtained from the FIREDOCK server at http://bioinfo3d.cs.tau.ac.il/FireDock/. All parameters were set as default, except that both proteins were treated as bound molecules.

FASTCONTACT rapidly estimates contact and binding free energies for protein−protein complexes. It is based on a statistically determined desolvation contact potential and Coulomb electrostatics with a distance-dependent dielectric contact. The scores were obtained from the FASTCONTACT server at http://structure.pitt.edu/servers/fastcontact/.

DFIRE uses an all-atom knowledge-based potential and provides accurate predictions as far as the stability of the complexes is concerned. Scores were obtained using the DFIRE Web server at http://sparks.informatics.iupui.edu/hzhou/dfire.html.

ZRANK is a high-performing scoring function that combines detailed electrostatics, van der Waals, and desolvation terms to calculate the energy of a complex. The scores were obtained by running the local version of the software.

Finally, a local version of ROSETTA (ROSETTA v2.1.2) was used to calculate the energy of each complex. Binding energies (ROSETTA score) were calculated by subtracting the energy of the complex from the energies of the individual chains. The I_int ROSETTA score is the component in the scoring function of the ROSETTADOCK algorithm that calculates a score for the interface in a given complex.

**Statistical Treatment of the Derived Data.** To find if there is a linear correlation between binding affinities and the various scores and their component, the Pearson product-moment correlation coefficient was calculated. It gives a parametric measure of correlation and assesses how well an arbitrary

monotonic function describes the relationship between two variables, without any assumptions about their frequency distribution. This correlation coefficient is obtained by dividing the covariance of the two variables by the product of their standard deviations. It has already been used successfully in protein−ligand docking for the assessment of the relationship between scoring functions and the binding affinity.[45−49] The coefficient ranges from −1 to 1: A value of 1 suggests that a linear correlation describes the relationship perfectly and positively; a value of 0 shows that there is no correlation between the two variables and a value of −1 describes a perfect, but negative, linear correlation (anticorrelation). The square of the Pearson product-moment correlation coefficient is used as a measure of the association between the two variables: for example, if $r^2 = 0.70$, then 70% of the variance of one variable can be explained by changes in the other variable and the linear relationship between them. Pearson product-momentum correlation coefficient and $p$-values were determined for each data set.

## Results

**The Protein−Protein Binding Affinity Benchmark.** We have compiled a protein−protein binding affinity benchmark that includes the experimentally characterized dissociation constants ($K_d$) of 81 protein−protein complexes along with the method used to determine them (Table 1). It also includes the experimental conditions (pH and temperature) at which the $K_d$ values were measured. Within this benchmark, the Chymotrypsin−Ecotin protein−inhibitor complex has the highest affinity with a value of $4.4 \times 10^{-15}$ M as determined by stopped-flow fluorimetry at 25 °C and 7.0 pH.[50] The lowest binding affinity complex ($K_d < 6.5 \times 10^{-4}$ M) is the one formed by the UEV domain and Ubiquitin as determined by surface plasmon resonance at 20 °C and 7.2 pH.[51] The reported affinities cover an extremely broad range of about 11 orders of magnitude. We further classified the complexes based on their binding affinity into three groups: high ($K_d \leq 10^{-9}$ M), medium ($10^{-9}$ M $< K_d \leq 10^{-6}$ M), and low ($K_d > 10^{-6}$ M) binding affinity complexes.

**Scoring of the Original Crystal Structures To Check for Bias in the Refinement Procedure.** Scoring of the initial crystal structures was performed in order to check for a potential bias in our results introduced by the energy minimization and force field used in HADDOCK. The crystal structures of the protein−protein binding affinity benchmark were scored directly, without any optimization and plotted against the affinity data (Supporting Information, Figure 1s, A). The resulting scores were compared directly to those of the minimized and the water-refined structures (Supporting Information, Figure 1s, B). These results indicate that the scoring functions tested in this study are not substantially affected by the treatment of the initial input structures. Scores calculated by ATTRACT, DFIRE, FASTCONTACT, FIREDOCK, PYDOCK and ROSETTADOCK use built-in structure refinement protocols and, thus, are the result of a reoptimization of the input structures. Very few structures have significantly different scores. PISA and AFFINITYSCORE1.0 use descriptors that do not change significantly during refinement (e.g., buried surface area, gap volume, total number of solvent exposed charged and hydrophobic atoms, etc.). When the complexes are optimized, an improvement in the correlation of affinities to AFFINITYSCORE1.0 is observed (Supporting Information, Table 1s, Compare Figure 1 to Supporting Information, Figures 1s and 2s). The various scores (original,

minimized, water-refined) show a very strong linear dependence (Supporting Information, Figure 1s). ZRANK was the only algorithm that showed a minor dependence on the treatment of the crystal structures (∼20% of the structures were calculated to have different ZRANK score, shown in Supporting Information Figure 1s), but the correlations to the experimentally determined binding affinities was not significantly altered (Supporting Infomation Table 1s).

**Performance of Scoring Functions: Poor Correlation with the Experimental $K_d$'s.** We have tested nine scoring functions implemented in well established docking programs or Web servers for their ability to correlate with the experimental binding affinities; these include ATTRACT,[38] DFIRE,[39] FASTCONTACT,[40] FIREDOCK,[41] HADDOCK,[36] PISA,[42] PYDOCK,[43] ROSETTADOCK[27] and ZRANK.[44] AFFINITYSCORE1.0,[21] a binding affinity prediction algorithm for protein−protein complexes, was also tested. All complexes were treated in a similar way to build missing atoms and remove possible clashes. They were either energy minimized or subjected to a gentle refinement in explicit water using the HADDOCK Web server (http://haddock.chem.uu.nl) (see Material and Methods).

All scoring algorithms failed to correlate the value of the binding affinity (p$K_d$) to the score that was calculated for each complex by the corresponding scoring function, both for water-refined (Figure 1) and energy minimized only complexes (Supporting Information, Figure 2s). All calculated $r$ values are between −0.32 and 0.18. The highest correlation ($r = -0.32$) was obtained with FIREDOCK for the water refined complexes (Figure 1G). This correlation is, however, still low and has very limited predictive power. In general, all scoring functions poorly correlate to the binding affinity and the corresponding $r$ values do not considerably change when complexes are refined in water (Figure 1 and Supporting Information, Figure 1s), except for AFFINITYSCORE1.0, HADDOCK and FASTCONTACT that show somewhat better (but still low) correlations after water refinement.

When the components of each scoring function are analyzed for their correlation to binding affinity (Supporting Information, Tables 2s and 3s), interesting observations can be made: the correlation for the hydrogen bonding function of FIREDOCK reaches almost −0.33 for the energy-minimized complexes (Supporting Information, Figure 2s) (one of the highest values among all analyzed components). After water refinement, this value drops to −0.28. The water refinement apparently rearranges the hydrogen bonding network of the molecules at the interface. An increased correlation after water refinement is detected for the van der Waals component of HADDOCK (from $r = -0.22$ to $r = -0.33$). Similar increases are observed for the FIREDOCK and ROSETTADOCK attractive van der Waals interactions (Supporting Information, Tables 2s and 3s).

Only few of the various score components (Supporting Information, Tables 2s and 3s) show statistically significant $r$ values: The HADDOCK van der Waals potential reaches a $p$-value of 0.0027 for the water-refined complexes ($r = -0.32$); the hydrogen bonding term of FIREDOCK reaches $p$-values of 0.0029 ($r = -0.33$) and 0.0125 ($r = -0.28$), for the complexes that were minimized and refined in water, respectively, and some knowledge-based components of the ROSETTADOCK score describing secondary structure energy terms exhibit significant $p$-values (e.g., 0.001 for the backbone−backbone hydrogen bond term ($r = -0.36$)) (Supporting Information, Tables 2s and 3s).

**Table 1.** The Binding Affinity Benchmark Including $K_d$ Values for 81 Protein–Protein Complexes

| | PDB ID (complex) | protein 1 (receptor) | protein 2 (ligand) | $K_d$ (mol/L)[a] | $pK_d$ | $\Delta G^{\theta b}$ (kcal/mol) | method[c] | pH | $T_{(oC)}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1N8O_ABC:E | Chymotrypsin | Ecotin | $4.4 \times 10^{-15}$ | 14.36 | −19.57 | a$^{FRET}$ | 8.0 | 25 |
| 2 | 7CEI_A:B | Colicin E7 nuclease | Im7 immunity protein | $5.0 \times 10^{-15}$ | 14.30 | −19.50 | f | 7.0 | 25 |
| 3 | 1DFJ_E:I | Ribonuclease A | Rnase inhibitor | $5.9 \times 10^{-14}$ | 13.23 | −18.04 | b | 6.0 | 25 |
| 4 | 1AVX_A:B | Porcine trypsin | Soybean trypsin inhibitor (Kunitz) | $6 \times 10^{-14}$ | 13.22 | −18.03 | c | 8.0 | 25 |
| 5 | 1AY7_A:B | Barnase | Barstar | $6 \times 10^{-14}$ | 13.22 | −18.03 | a | 8.0 | 25 |
| 6 | 1BVN_P:T | α-amylase | Tendamistat | $9 \times 10^{-12}$ | 11.05 | −15.06 | d | 7.0 | 25 |
| 7 | 1IQD_AB:C | Fab | Factor VIII domain C2 | $1.4 \times 10^{-11}$ | 10.85 | −14.80 | e | 5.0 | 25 |
| 8 | 1CGI_E:I | Chymothripsinogen A | PSTI3 | $1.6 \times 10^{-11}$ | 10.80 | −14.72 | e | 8.0 | 25 |
| 9 | 1MAH_A:F | Acetylcholinesterase | Fasciculin 2 | $2.5 \times 10^{-11}$ | 10.60 | −14.45 | c | 7.5 | 25 |
| 10 | 1XQS_A:C | HspBP1 | Hsp70 ATPase domain | $4.0 \times 10^{-11}$ | 10.40 | −14.18 | f | 7.2 | 25 |
| 11 | 1EZU_C:AB | D102N Trypsin | Y69F D70P Ecotin | $8.0 \times 10^{-11}$ | 10.10 | −13.77 | e | 8.0 | 25 |
| 12 | 1JPS_HL:T | Fab D3H44 | Tissue factor | $1.0 \times 10^{-10}$ | 10.00 | −13.63 | d | 7.4 | 25 |
| 13 | 1PPE_E:I | Bovine trypsin | CMTI-1 squash inhibitor | $1.0 \times 10^{-10}$ | 10.00 | −13.50 | g | 8.3 | 22 |
| 14 | 1IBR_A:B | Ran GTPase | Importin β | $3.0 \times 10^{-10}$ | 9.52 | −12.98 | c | 7.4 | 25 |
| 15 | 1R0R_A:C | Subtilisin Carlsberg | OMTKY | $3.4 \times 10^{-10}$ | 9.47 | −12.74 | f | 8.3 | 21 |
| 16 | 1T6B_X:Y | Anthrax protective antigen | Anthrax toxin receptor | $4.0 \times 10^{-10}$ | 9.40 | −12.81 | d | 7.4 | 25 |
| 17 | 1KXP_A:D | Actin | Vitamin D binding protein | $1.0 \times 10^{-9}$ | 9.0 | −12.27 | c | 7.5 | 25 |
| 18 | 2FD6_HL:U | Urokinase Plasminogen receptor antibody | Urokinase Plasminogen activator receptor | $1.0 \times 10^{-9}$ | 9.0 | −12.19 | c | 7.4 | 23 |
| 19 | 2I25_N:L | Shark single domain antigen receptor | Lysozyme | $1.0 \times 10^{-9}$ | 9.0 | −12.27 | d | 7.4 | 25 |
| 20 | 2VIS_AB:C | Fab | Flu virus hemagglutinin | $1.0 \times 10^{-9}$ | 9.0 | −12.27 | d | 7.4 | 25 |
| 21 | 1BGX_HL:T | Fab | Taq polymerase | $1.0 \times 10^{-9}$ | 9.0 | −12.27 | d | 7.4 | 25 |
| 22 | 2B42_A:B | Xylanase | Xylanase inhibitor | $1.07 \times 10^{-9}$ | 8.97 | −12.11 | d | 5.0 | 22 |
| 23 | 1EAW_A:B | Matriptase | BPTI | $1.7 \times 10^{-9}$ | 8.77 | −11.96 | e | 7.5 | 25 |
| 24 | 2JEL_HL:P | Fab Jel42 | HPr | $2.8 \times 10^{-9}$ | 8.55 | −11.58 | a$^{Polarization}$ | 7.2 | 23 |
| 25 | 1ML0_AB:D | Viral chemokine binding protein M3 | Chemokine Mcp1 | $3.0 \times 10^{-9}$ | 8.52 | −11.62 | d | 7.4 | 25 |
| 26 | 1KKL_ABC:H | HPr kinase C-ter domain | HPr | $3.1 \times 10^{-9}$ | 8.51 | −11.60 | d | 8.0 | 25 |
| 27 | 1BKD_R:S | Ras GTPase | Son of Sevenless | $3.3 \times 10^{-9}$ | 8.48 | −11.56 | d | 7.4 | 25 |
| 28 | 1BJ1_HL:VW | Fab | vEGF | $3.4 \times 10^{-9}$ | 8.47 | −11.55 | d | 7.4 | 25 |
| 29 | 1KXQ_H:A | Camel VHH | Pancreatic α-amylase | $3.5 \times 10^{-9}$ | 8.46 | −11.53 | d$^{IAsys}$ | 7.4 | 25 |
| 30 | 1EWY_A:C | Ferredoxin reductase | Ferredoxin | $3.57 \times 10^{-9}$ | 8.45 | −11.52 | a | 8.0 | 25 |
| 31 | 1KAC_A:B | Adenovirus fiber knob protein | Adenovirus receptor | $4.75 \times 10^{-9}$ | 8.32 | −11.35 | d | 7.4 | 25 |
| 32 | 1OPH_A:B | α−1-antitrypsin | Trypsinogen | $5 \times 10^{-9}$ | 8.30 | −11.32 | a$^{FRET}$ | 7.4 | 25 |
| 33 | 1M10_A:B | Von Willebrand Factor Domain A1 | Glycoprotein IB-α | $5.8 \times 10^{-9}$ | 8.24 | −11.23 | d | 7.4 | 25 |
| 34 | 1Y64_A:B | Actin | BN1 protein | $8.1 \times 10^{-9}$ | 8.09 | −11.03 | a | 7.0 | 25 |
| 35 | 1TMQ_A:B | α-amylase | RAGI inhibitor | $1.1 \times 10^{-8}$ | 7.96 | −10.85 | e | 8.0 | 25 |
| 36 | 2AJF_A:E | ACE2 | SARS spike protein receptor binding domain | $1.62 \times 10^{-8}$ | 7.79 | −10.62 | d | 7.4 | 25 |
| 37 | 1IJK_A:BC | Von Willebrand Factor Domain A1 | Botrocetin | $2.3 \times 10^{-8}$ | 7.64 | −10.41 | c | 7.4 | 25 |
| 38 | 1H1V_A:G | Actin | Gelsonin | $2.3 \times 10^{-8}$ | 7.64 | −10.24 | b | 7.0 | 20 |
| 39 | 1DE4_AB:CF | HFE | Transferrin receptor ectodomain | $2.4 \times 10^{-8}$ | 7.62 | −10.39 | d | 7.4 | 25 |
| 40 | 1E6J_HL:P | Fab | HIV-1 capsid protein 24 | $2.9 \times 10^{-8}$ | 7.53 | −10.28 | d | 7.4 | 25 |
| 41 | 2HLE_A:B | Ephrin B4 receptor | Ephrin B2 ectodomain | $4.0 \times 10^{-8}$ | 7.4 | −10.09 | f | 7.8 | 25 |
| 42 | 1ATN_A:D | Actin | Dnase I | $4.5 \times 10^{-8}$ | 7.35 | −10.02 | e | 8.4 | 25 |
| 43 | 1WEJ_HL:F | Fab E8 | Cytochrome C | $6 \times 10^{-8}$ | 7.22 | −9.85 | f | 7.2 | 25 |
| 44 | 1A2K_C:AB | Ran GTPase | Nuclear Transport Factor 2 | $1 \times 10^{-7}$ | 7.00 | −9.54 | f | 7.5 | 25 |
| 45 | 2C0L_A:B | PTS1 and TRP region of PEX5 | SCP2 | $1.09 \times 10^{-7}$ | 6.96 | −9.81 | f | 7.4 | 35 |
| 46 | 1RLB_ABCD:E | Transthyretin | Retinol binding protein | $1.34 \times 10^{-7}$ | 6.87 | −9.37 | a | 7.4 | 25 |
| 47 | 1GLA_G:F | Glycerol kinase | Glucose specific phosphocarrier | $1.8 \times 10^{-7}$ | 6.74 | −9.20 | d | 7.4 | 25 |
| 48 | 2HRK_A:B | Glutamyl-t-RNA synthetase | GU-4 nucleic binding protein | $1.93 \times 10^{-7}$ | 6.71 | −9.15 | d | 7.4 | 25 |
| 49 | 1GP2_A:BG | Gi-α | Gi-β,γ | $2.54 \times 10^{-7}$ | 6.60 | −9.14 | c | 7.5 | 30 |
| 50 | 1FAK_HL:T | Coagulation factor VIIa | Soluble tissue factor | $2.8 \times 10^{-7}$ | 6.55 | −8.93 | d | 7.5 | 25 |
| 51 | 1GRN_A:B | CDC42 GTPase | CDC42 GAP | $3.88 \times 10^{-7}$ | 6.41 | −8.74 | a | 8.0 | 25 |
| 52 | 1E6E_A:B | Adrenoxin reductase | Adrenoxin | $0.86 \times 10^{-6}$ | 6.07 | −8.27 | d | 7.4 | 25 |
| 53 | 1J2J_A:B | Arf1 GTPase | GAT domain of GGA1 | $1.1 \times 10^{-6}$ | 5.96 | −8.12 | d | 8.0 | 25 |
| 54 | 1BUH_A:B | CDK2 Kinase | Ckshs1 | $1.5 \times 10^{-6}$ | 5.82 | −7.54 | f | 7.5 | 10 |
| 55 | 1GPW_A:B | HISF protein | Amidotransferase 6 | $1.5 \times 10^{-6}$ | 5.82 | −7.94 | e | 8.5 | 25 |
| 56 | 2OT3_B:A | Rab21 GTPase | Rabex-5 VPS9 domain | $1.8 \times 10^{-6}$ | 5.74 | −7.83 | a$^{Polarization}$ | 7.4 | 25 |
| 57 | 2BTF_A:P | Actin | Profilin | $2.3 \times 10^{-6}$ | 5.70 | −7.69 | a | 7.0 | 25 |
| 58 | 1EFN_B:A | HIV1-NEF protein | SH3 domain | $2.5 \times 10^{-6}$ | 5.60 | −7.64 | d | 7.4 | 25 |
| 59 | 1HE8_B:A | Ras GTPase | PIP3 kinase | $2.5 \times 10^{-6}$ | 5.60 | −7.51 | a$^{Polarization}$ | 7.5 | 20 |
| 60 | 1B6C_A:B | FKBP Binding Protein | TGFβ receptor | $2.8 \times 10^{-6}$ | 5.55 | −7.57 | d | 7.4 | 25 |
| 61 | 1SBB_A:B | T-cell receptor β | Staphylococcus enterotoxin B | $3.00 \times 10^{-6}$ | 5.52 | −7.53 | d | 7.5 | 25 |
| 62 | 1I4D_D:AB | Rac GTPase | Arfaptin | $3 \times 10^{-6}$ | 5.52 | −7.45 | f | 8.7 | 22 |
| 63 | 1GHQ_A:B | Complement C3 | Epstein−Barr visus receptor CR2 | $4.3 \times 10^{-6}$ | 5.37 | −7.32 | d | 7.4 | 25 |
| 64 | 2MTA_HL:A | Methylamine Dehydrogenase | Amicyanin | $4.5 \times 10^{-6}$ | 5.35 | −7.29 | g | 7.5 | 25 |
| 65 | 1E96_A:B | Rac GTPase | p67 Phox | $6 \times 10^{-6}$ | 5.22 | −6.95 | f | 7.0 | 18 |
| 66 | 1Z0K_A:B | Rab4A GTPase | RAB4 binding domain of Rabenosyn | $7.2 \times 10^{-6}$ | 5.14 | −7.01 | d | 7.5 | 25 |
| 67 | 1QA9_A:B | CD2 | CD58 | $9 \times 10^{-6}$ | 5.05 | −7.16 | d | 7.4 | 37 |

**Table 1.** Continued

| | PDB ID (complex) | protein 1 (receptor) | protein 2 (ligand) | $K_d$ (mol/L)[a] | $pK_d$ | $\Delta G^{\theta b}$ (kcal/mol) | method[c] | pH | $T_{(oC)}$ |
|---|---|---|---|---|---|---|---|---|---|
| 68 | 1FC2_C:D | Staphylococcus protein A | Human Fc fragment | $1 \times 10^{-5}$ | 5.00 | −6.82 | d | 7.2 | 25 |
| 69 | 1AZS_AB:C | Adenylyl cyclase | AC activator Gs α complex | $1 \times 10^{-5}$ | 5.00 | −6.82 | e$^{Scintillation}$ | 8.0 | 25 |
| 70 | 2PCC_A:B | Cyt C peroxidase | Cytochrome C | $1 \times 10^{-5}$ | 5.00 | −6.82 | f | 6.0 | 25 |
| 71 | 1AK4_A:D | Cyclophilin | HIV capsid | $1.60 \times 10^{-5}$ | 4.80 | −6.43 | f | 6.5 | 20 |
| 72 | 1GCQ_B:C | GRB2 C-ter SH3 domain | GRB2 N-ter SH3 domain | $1.68 \times 10^{-5}$ | 4.77 | −6.51 | d | 7.4 | 25 |
| 73 | 1WQ1_R:G | Ras GTPase | Ras GAP | $1.7 \times 10^{-5}$ | 4.77 | −6.50 | b | 7.5 | 25 |
| 74 | 1I2M_A:B | Ran GTPase | RCC1 | $3.16 \times 10^{-5}$ | 4.50 | −6.14 | b | 7.4 | 25 |
| 75 | 2OOB_A:B | Ubiquitin ligase | Ubiquitin | $6.0 \times 10^{-5}$ | 4.22 | −5.76 | f | 7.0 | 25 |
| 76 | 1NW9_B:A | Caspase-9 | XIAP-BIR3 | $7.4 \times 10^{-5}$ | 4.13 | −5.73 | a | 6.5 | 30 |
| 77 | 1AKJ_AB:DE | MHC Class I HLA-A2 | T-cell CD8 coreceptor | $1.26 \times 10^{-4}$ | 3.90 | −5.32 | d | 7.4 | 25 |
| 78 | 1R8S_A:E | Arf1 GTPase | Sec 7 domain | $1.50 \times 10^{-4}$ | 3.82 | −5.21 | a$^{Polarization}$ | 8.0 | 25 |
| 79 | 1IB1_AB:E | 14−3−3 protein | Serotonin N-acetylase | $2.34 \times 10^{-4}$ | 3.63 | −4.87 | f | 7.0 | 20 |
| 80 | 2HQS_A:H | TolB | Pal | $4.70 \times 10^{-4}$ | 3.33 | −4.46 | f | 7.5 | 20 |
| 81 | 1S1Q_A:B | UEV domain | Ubiquitin | $6.35 \times 10^{-4}$ | 3.19 | −4.29 | d | 7.2 | 20 |

[a] Corresponding references for each biochemically determined binding affinity value are included in Supporting Information, Detailed materials and methods. [b] $\Delta G^{\theta}$ corresponds to the calculated Gibbs free-energy for the system, estimated directly from $\Delta G^{\theta} = -RT \ln K_d$, where $R$ is the gas constant and $T$ the absolute temperature. $K_d$ corresponds to the experimentally determined binding affinity, $K_d = k_{on}/k_{off}$. [c] Methods: (a) fluorescence spectroscopy; (b) stopped-flow fluorimetry; (c) radioligand binding (competitive binding experiments); (d) surface plasmon resonance; (e) spectrophotometric assays; (f) isothermal titration calorimetry; (g) HPLC/UV absorption analysis.

Even the affinity-based predictor AFFINITYSCORE 1.0 had limited predictive power over the whole data set. Although its score shows one of the highest correlations ($r = -0.29$) with experimental data (Figure 1), it is still low. However, the quality of the predictions is greatly improved when the complexes are refined in water compared to the direct scoring of the original crystal structures or of the minimized complexes (Figure 1 and Supporting Information). Interestingly, grouping the experimental data into methodology-oriented categories reduces the prediction noise: Most of the algorithms, including AFFINITY-SCORE1.0, now show significant correlation coefficients to several experimental techniques (see below).

**Weak Correlations Emerge When Experimental Binding Affinities Are Grouped According to the Method by Which They Were Determined.** If complexes are classified according to the method by which their dissociation constants have been determined, significant correlations emerge (see Figure 1, Table 2 and Supporting Information). AFFINITYSCORE1.0 exhibits the highest correlation to the data accumulated from spectrophotometric assays, with an $R^2$ reaching 0.55. Other algorithms perform equally well in predicting this data set, as shown in Table 2. However, due to the limited amount of data (only 9 complexes were found to have a dissociation constant determined by spectrophotometric assays), these results should be treated with caution. For example, although DFIRE score strongly correlates to these data ($r = 0.78$), this correlation is positive, meaning that the DFIRE predictions and the experimental affinities are anticorrelated for this particular data set!

When more measurements are considered from standardized techniques, surface plasmon resonance (SPR) data appear to correlate weakly to the DFIRE score (Table 2 and Figure 1), the PISA entropy component ($T\Delta S$) and the FIREDOCK and PYDOCK scores. PYDOCK, FASTCONTACT and AFFINITY-SCORE1.0 show significant correlation coefficients when compared to data from Isothermal Titration Calorimetry experiments. However, the scoring performance of the various algorithms tends to be orthogonal when considering binding measurements from the other techniques (fluorescence- and radioligand-based techniques).

**Linear Correlation Emerges When Grouping the Low, Medium, and High Affinity Complexes.** Considering the poor performance of scoring functions in predicting individual binding affinities, we investigated if the scores would have a predictive power in classifying the complexes in three categories: high, medium, or low affinity according to the classification in Table 1. For this, the average binding affinity and scores (with their corresponding standard deviations) were calculated for each of the three classes of complexes. Their putative interdependence was assesed by the Pearson-product momentum coefficient. Results show that high correlations emerge: Five out of ten algorithms have at least one component that correlates nicely to the binding affinity (Table 3). All components are shown in Supporting Information Tables Table 2s and 3s. Statistically significant correlations are found for the buried surface area (present is several scoring functions), the HADDOCK score and its van der Waals component (after water refinement), the FIREDOCK score and its attractive van der Waals component and several components of the ROSETTA-DOCK score (Table 3). Note that several components only show significant correlation to binding affinity after water-refinement of the complexes (Table 3, B Rows). Another interesting observation is that the only solvation term that correlates with binding affinity is the Lazaridis-Karplus desolvation energy term,[52] implemented in the ROSETTADOCK scoring function.

When the refined complexes are divided into four categories instead of three (High ($K_d < 10^{-9}$ M), Medium ($10^{-9}$ M $\leq K_d$ <$10^{-7}$ M), Medium/Low ($10^{-7}$ M $\leq K_d$ <$10^{-5}$ M), Low/Very Low ($K_d \geq 10^{-5}$ M)), several scores still show robust correlations:

• The FIREDOCK score ($r = -0.96$) along with two of its components, the London dispersion forces ($r = 0.97$) and Hydrogen bonds ($r = -0.89$)

• The ROSETTADOCK London dispersion forces ($r = 0.94$) and the backbone−backbone hydrogen bonding term corresponding to distant residues in the primary sequence of the complexes ($r = -0.99$)

• The Electrostatics component of the FASTCONTACT score ($r = -0.96$).

Even a further categorization of the binding affinities into 5 classes (Very High ($K_d < 10^{-10}$ M), High ($10^{-10}$ M $\leq K_d < 10^{-8}$ M), Medium ($10^{-8}$ M $\leq K_d$ <$10^{-6}$ M), Low ($10^{-6}$ M $\leq K_d$ <$10^{-5}$ M), Very Low ($K_d \geq 10^{-5}$ M)) reveals again some significant correlations. The FireDock score and FIREDOCK Hydrogen bonds have still high correlations to the affinity-based groups, with $r$'s of $-0.92$. Some other correlations are also constant,
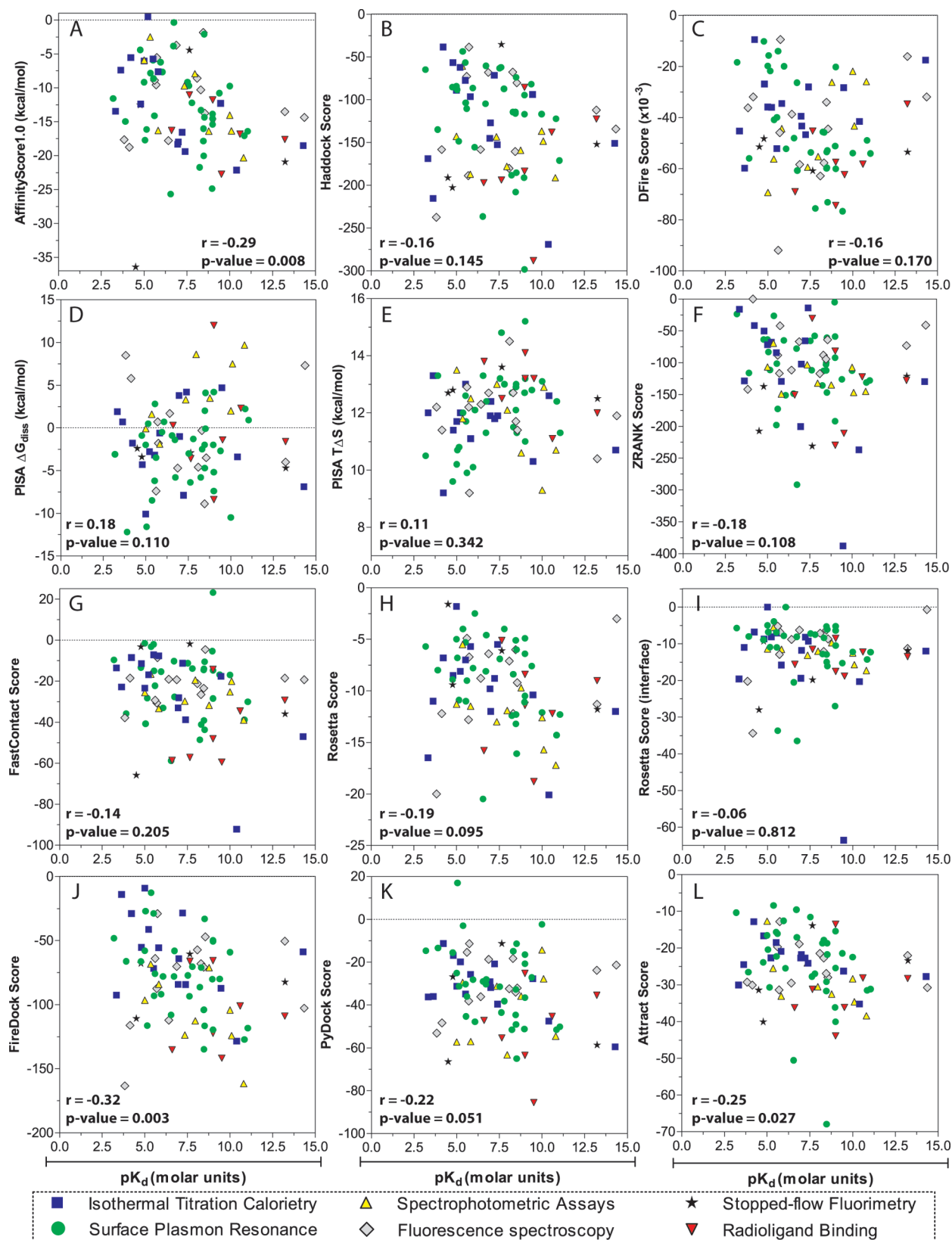
**Figure 1.** Scatter plots of binding affinity ($-\log(K_d)$) (*x*-axis) versus score (*y*-axis) for the 81 water refined complexes from our binding affinity benchmark. The scores were calculated using nine different scoring functions and a binding affinity prediction algorithm (panels A−L, see text). *r* and *p*-values are indicated in each plot. Different colors and shapes of the data points correspond to different methodology followed to experimentally determine binding affinity.

including the FASTCONTACT electrostatics ($r = -0.88$), the ROSETTA score ($r = -0.88$) and its Electrostatics component

($r = -0.95$). An analysis of the minimized complexes reveals similar trends (data not shown). Another interesting observa-

**Table 2.** Pearson's Correlation Coefficients between Scoring of the Water Refined Complexes and Experimentally Determined Dissociation Costants ($K_d$'s)[a]

|  | binding affinity data (N = 81) | surface plasmon resonance (N = 33) | isothermal titration calorimetry (N = 15) | fluorescence spectroscopy and FRET (N = 13) | spectrophotometric assays (N = 9) | radioligand binding (N = 7) | stopped-flow fluorimetry (N = 4) |
|---|---|---|---|---|---|---|---|
| AFFINITY1.0 | **−0.29** | −0.29 | **−0.61** | 0.04 | **−0.74** | −0.36 | 0.15 |
| HADDOCK | −0.16 | *−0.34* | −0.31 | 0.19 | −0.40 | 0.35 | 0.32 |
| PISA (ΔG_diss) | 0.18 | 0.24 | −0.13 | −0.10 | **0.73** | 0.01 | −0.12 |
| PISA (TΔS) | 0.11 | **0.42** | −0.19 | −0.11 | *−0.59* | −0.60 | −0.54 |
| ZRANK | −0.18 | −0.12 | *−0.50* | 0.20 | −0.51 | −0.10 | 0.58 |
| FIREDOCK | **−0.32** | **−0.42** | −0.42 | 0.32 | *−0.63* | −0.02 | −0.20 |
| FASTCONTACT | −0.14 | −0.08 | **−0.63** | 0.30 | −0.28 | 0.54 | −0.02 |
| ROSETTA | *−0.19* | *−0.34* | −0.33 | 0.44 | **−0.71** | 0.13 | −0.670 |
| ROSETTA[b] | −0.06 | −0.15 | −0.33 | *0.53* | **−0.71** | 0.061 | 0.447 |
| DFIRE | −0.16 | **−0.58** | 0.33 | 0.24 | **0.78** | 0.59 | 0.38 |
| PYDOCK | **−0.22** | **−0.38** | *−0.66* | 0.41 | 0.32 | 0.214 | −0.09 |
| ATTRACT | **−0.25** | *−0.33* | *−0.50* | −0.05 | **−0.70** | 0.216 | 0.39 |

[a] Numbers highlighted in both bold and italics show very significant correlations ($p$-value <0.01), bold type shows significant correlations ($0.01 \leq p$-value <0.05), whereas numbers highlighted in italics show correlations of potential interest ($0.05 \leq p$-value <0.1). [b] Rosetta scoring of the interface (I_int).

tion is that HADDOCK Electrostatics also shows very high $r$'s when classified into 4 ($r = −0.96$) or 5 ($r = −0.83$) categories but not into 3 ($r = −0.89$, not statistically significant). The electrostatics component of ROSETTADOCK is also robust when the complexes are classified in 4 ($r = −0.91$) or 5 ($r = −0.89$) categories.

These results indicate that several of the algorithms tested in this work incorporate a component that might be meaningful and useful for a qualitative description of binding affinity; some of them remain robust when defining more than three classes. Unfortunately, all terms are associated with huge standard deviations (see Table 3), underlying that neither the algorithms nor their components alone can in reality predict if a given complex is of high, medium or low affinity. For that reason, the presented results should be handled with caution.

## Discussion

This work represents the first systematic assessment of the performance of scoring functions in protein−protein docking for the prediction of binding affinities. Our binding affinity benchmark comprises $K_d$ values described in the literature for 81 complexes from the Protein−Protein docking Benchmark 3.0.[28] Our results clearly demonstrate that current scoring functions are very much limited in their ability to predict binding affinity.

One reason is that scoring in biomolecular docking has been developed to discriminate the native structure from thousands of decoy structures, and not for binding affinity prediction since we typically already know in protein−protein docking that a given pair will interact. However, the two problems need to converge, in order to have reliable scoring functions that discriminate false positives along with correctly predicted structures. An important conclusion from this work is that, given the structure of three proteins, and knowing *a priori* that two of them bind and one does not, we are still very much limited in our ability to predict the correct pair.

Moreover, binding affinity greatly depends on temperature, pH and salt concentration[53] and can also be affected by high concentrations of other macromolecules, a phenomenon that is described as macromolecular crowding.[54] Thus, it seems highly unlikely to directly correlate a function, even knowledge-based, to experimental, biochemically determined $K_d$'s: from experiment to experiment, temperature, pH and salt concen-

trations might change and our benchmark, which is the first of its kind, might well reflect that. Considering the results from Figure 1 and Table 2 (as well as the Supporting Information) it is evident that scoring functions exhibit weak, but significant correlations to binding affinities determined by Surface Plasmon Resonance, Isothermal Titration Calorimetry and various Spectrophotometric Assays. On the other hand, algorithms were completely orthogonal when fluorescence data were put to the test. Noise in experimental data can be an additional factor that influences the performance of the various algorithms in predicting the affinity. The very poor correlations observed with fluorescence-based techniques, such as resonance energy transfer (FRET), fluorescence polarization or anisotropy with tagged molecules might indeed correspond to the inherent limitations of these techniques (e.g., general problems of fluorophores). Binding affinity data measured through Radio-ligand binding assays are also orthogonal to scoring. These results might stem from the fact that results obtained by these methods are highly influenced by slight variations in the experimental procedures. On the other hand, the label-free SPR technique and thermodynamic measurements with ITC along with traditional methods for affinity measurements give better correlations between the experimental data and the scores calculated by the algorithms; still, these correlations remain low.

It is known that the mechanism of protein binding is highly dependent on the particular system:[55] Depending on the ion concentration, the association rate of fast binders ($k_{on} > 108$ $M^{−1}$ $s^{−1}$) might significantly change[56] in comparison to the corresponding value of slow binders ($k_{on} < 108$ $M^{−1}$ $s^{−1}$) that remain much more constant.[57] The fast binding mechanism is based on long-range electrostatic interactions that drive the molecules to find each other,[58] whereas the slow binding process is mostly controlled by desolvation forces and van der Waals interactions.[57] The fact that a really high $k_{on}$ does not necessarily reflect a $K_d$ that can possibly describe a high affinity complex is overseen by the algorithms.

Scoring algorithms need to incorporate biochemical information that relates to the $k_{on}$ value and possibly give different weights to the Electrostatic or the van der Waals components. This can be achieved by analyzing structural parameters of the input 3D structures. For example, high-throughput comparative research aiming at deciphering global structural features of fast

**Table 3.** Statistical Analysis of the Various Scoring Functions[a]

| scoring function | component | A/B | average high | average medium | average low | r value, p-value |
|---|---|---|---|---|---|---|
| HADDOCK | HADDOCK score | A | **−146.4 ± 49.4** | **−130.2 ± 45.2** | **−111.0 ± 47.6** | **−0.998, 0.038** |
| | | B | **−152.9 ± 64.6** | **−131.6 ± 54.1** | **−114.7 ± 57.9** | **−0.998, 0.037** |
| | van der Waals | B | **−82.5 ± 33.4** | **−69.3 ± 20.0** | **−54.8 ± 24.9** | **−0.999, 0.021** |
| | Buried Surface Area | A | **2340.2 ± 1163.3** | **2043.2 ± 561.4** | **1757.5 ± 660.4** | **∼1.000, 0.001** |
| | | B | *2345.1 ± 1071.8* | *2133.3 ± 589.7* | *1789.6 ± 707.1* | *0.989, 0.093* |
| PISA | *PISA ΔG (dissociation)* | *A* | *0.0 ± 5.7* | *−1.3 ± 4.1* | *−2.1 ± 5.3* | *0.988, 0.097* |
| | | *B* | *0.2 ± 6.2* | *−1.3 ± 4.0* | *−2.4 ± 4.7* | *0.997, 0.051* |
| | *Buried Surface Area* | *A* | **1118.8 ± 537.6** | **990.3 ± 268.5** | **857.8 ± 316.8** | **∼1.000, 0.012** |
| | | *B* | *1120.5 ± 538.8* | *990.3 ± 267.1* | *810.3 ± 353.4* | *0.995, 0.065* |
| FIREDOCK | FIREDOCK score | A | **−101.4 ± 32.6** | **−84.5 ± 26.7** | **−69.6 ± 36.6** | **∼−1.000, 0.017** |
| | | B | **−99.3 ± 32.5** | **−82.2 ± 26.9** | **−66.7 ± 35.3** | **∼−1.000, 0.011** |
| | attractive van der Waals | A | **−56.3 ± 17.8** | **−50.6 ± 12.0** | **−44.7 ± 15.6** | **∼−1.000, 0.010** |
| | | B | **−56.8 ± 18.0** | **−50.1 ± 12.6** | **−43.2 ± 15.6** | **∼−1.000, 0.014** |
| | *repulsive van der Waals* | *B* | *16.8 ± 9.1* | *14.7 ± 5.9* | *11.7 ± 7.0* | *0.993, 0.074* |
| | *analytical continuum electrostatics potential* | *A* | *−2.7 ± 16.2* | *−2.1 ± 10.2* | *−1.3 ± 12.1* | *−0.996, 0.055* |
| | hydrogen bonds | A | **−9.5 ± 3.7** | **−7.4 ± 3.4** | **−5.5 ± 3.2** | **∼−1.000, 0.013** |
| ROSETTA | strand−strand pairing | A | **−1.7 ± 3.0** | **−1.1 ± 2.8** | **−0.5 ± 1.4** | **−0.999, 0.020** |
| | residue pair potentials | A | *−10.9 ± 12.1* | *−9.6 ± 8.1* | *−7.9 ± 9.9* | *−0.995, 0.065* |
| | | B | **−10.9 ± 12.1** | **−9.6 ± 8.1** | **−8.2 ± 9.9** | **∼−1.000, 0.014** |
| | *van der Waals* | *A* | *0.2 ± 0.4* | *0.3 ± 0.5* | *0.5 ± 0.9* | *−0.994, 0.067* |
| | | *B* | *0.2 ± 0.4* | *0.3 ± 0.5* | *0.5 ± 0.8* | *−0.990, 0.091* |
| | $C_\beta$ packing density | A | **−2.0 ± 2.0** | **−2.1 ± 1.3** | **−2.3 ± 1.2** | **∼1.000, 0.011** |
| | | B | *−2.0 ± 2.0* | *−2.2 ± 1.3* | *−2.3 ± 1.2* | *0.993, 0.075* |
| | *strand pair distance/register* | *B* | *−0.6 ± 1.2* | *−0.5 ± 1.1* | *−0.3 ± 0.8* | *−0.993, 0.074* |
| | *attractive van der Waals potential* | *B* | *−48.2 ± 16.7* | *−39.6 ± 11.8* | *−33.4 ± 12.8* | *−0.996, 0.056* |
| | repulsive van der Waals | A | *15.9 ± 37.7* | *10.0 ± 14.2* | *5.8 ± 6.1* | *0.996, 0.056* |
| | Lazaridis-Karplus solvation | A | *32.5 ± 14.7* | *26.9 ± 9.6* | *23.1 ± 10.8* | *0.995, 0.061* |
| | | B | **32.5 ± 14.9** | **26.8 ± 9.4** | **22.2 ± 11.0** | **0.999, 0.031** |
| | side chain hydrogen bonds | B | **−7.9 ± 4.1** | **−6.4 ± 3.6** | **−5.2 ± 3.5** | **−0.999, 0.030** |
| | surface-area based solvation model | A | **−28.6 ± 15.2** | **−25.8 ± 6.7** | **−22.4 ± 8.1** | **−0.998, 0.037** |
| | *statistics based pair term, favors salt bridges* | *B* | *−3.6 ± 3.0* | *−3.4 ± 2.1* | *−3.0 ± 2.0* | *−0.992, 0.083* |
| | Buried surface area | A | **2326.4 ± 1140.3** | **2051.2 ± 564.2** | **1797.4 ± 646.3** | **∼1.000, 0.009** |
| | | B | **2327.2 ± 1135.1** | **2050.9 ± 561.5** | **1724.5 ± 656.6** | **0.998, 0.037** |
| | *Coulomb term of interface atoms* | *B* | *−78.1 ± 163.7* | *−74.8 ± 84.0* | *−69.2 ± 95.2* | *−0.988, 0.099* |
| AFFINITYSCORE1.0 | *ΔG[θ]* | *B* | *−15.9 ± 3.6* | *−12.7 ± 6.8* | *−10.7 ± 6.8* | *−0.995, 0.064* |

[a] The first and the second columns show the algorithms tested and each of its corresponding components. The third column indicates which complexes were analyzed: (A) energy minimized and (B) water refined (see Material and Methods). The next three columns illustrate the average of each affinity-related group, along with the standard deviation. The r-value and the p-value are related to the average binding affinity of each group, shown in the last two columns of the Table. **Bold** and *italics* numericals correspond to two groups of significant values, $0 <$ **p-value** $\leq$ **0.05** and *0.1 > p-value >0.05*, respectively. (All scores and their components are reported in Tables 2s and 3s in the Supporting information).

and slow macromolecular binders should greatly improve our knowledge in this field and help us develop new algorithms.

Despite the challenging aspect of the problem (binding affinity is related to free energy, one of the most challenging quantities to predict in computational studies), there is still some hope: The major outcome of this study is that several components of the scoring functions that were tested in this study correlate to affinity, when the latter are grouped in three sets (high, medium and low) and remain robust when classified in four, or even five categories. The predictive power is however limited due to the huge associated standard deviations. Despite this, our results point to new directions, where, in order to design novel algorithms that can successfully describe the binding affinity of a protein−protein complex, consensus functions, possibly combining scores or their components from different algorithms, need to be developed. These can also include classifiers trained by supervised learning. In protein−ligand drug design, nonlinear models, including, for example, Support Vector Machines,[59] have already been applied to this problem. Our benchmark will be a valuable catalyst to achieve this goal. It provides a starting point for the design of novel algorithms that might open the route to the prediction, characterization and annotation of complete interactomes by high-throughput computational docking.

**Supporting Information Available:** Detailed description of the material and methods including the protocol to refine and score the protein−protein complexes using HAD-DOCK, a description of the way the ROSETTA score was calculated and all data-mined references for the experimentally determined binding affinities used in this research; two additional figures showing the dependency of the calculated scores on the refinement process (Figure 1s) and the scatter plot of the calculated scores versus experimental binding affinities for the energy minimized structures (Figure 2s); three tables giving the correlation coefficients and their corresponding significance for all scoring functions (Table 1s) and full statistical analysis of the scores and their components for the energy minimized and water refined complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Pawson, T.; Nash, P. Protein-protein interactions define specificity in signal transduction. *Genes Dev.* **2000**, *14* (9), 1027–47.

(2) Wilson, I. A.; Stanfield, R. L. Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.* **1994**, *4* (6), 857–67.

(3) Jones, S.; Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (1), 13–20.

(4) Moffat, K.; Chait, B. T. Biophysical methods: doing more with less. *Curr. Opin. Struct. Biol.* **2003**, *13* (5), 535–7.

(5) Sprinzak, E.; Sattath, S.; Margalit, H. How reliable are experimental protein-protein interaction data. *J. Mol. Biol.* **2003**, *327* (5), 919–23.

(6) Service, R. STRUCTURAL BIOLOGY: Structural Genomics, Round 2. *Science* **2005**, *307* (5715), 1554–1558.

(7) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–73.

(8) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47* (4), 409–43.

(9) Lensink, M. F.; Mendez, R.; Wodak, S. J. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* **2007**, *69* (4), 704–18.

(10) Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. Assessment of CAPRI predictions in rounds 3−5 shows progress in docking procedures. *Proteins* **2005**, *60* (2), 150–69.

(11) Aloy, P.; Russell, R. B. Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **2006**, *7* (3), 188–97.

(12) Dell'Orco, D. Fast predictions of thermodynamics and kinetics of protein-protein recognition from structures: from molecular design to systems biology. *Mol. BioSyst.* **2009**, *5* (4), 323–34.

(13) Abagyan, D.; Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **1994**, *235* (3), 983–1002.

(14) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **2002**, *93* (7), 2395–2417.

(15) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268* (5214), 1144–9.

(16) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. a fast analytical method for the calculation of approximate born radii. *J. Phys. Chem. A* **1997**, *101* (16), 3005–3014.

(17) Schaefer, M.; Karplus, M. A Comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **1996**, *100* (5), 1578–1599.

(18) Baker, B. M.; Murphy, K. P. Prediction of binding energetics from structure using empirical parameterization. *Methods Enzymol.* **1998**, *295*, 294–315.

(19) Horton, N.; Lewis, M. Calculation of the free energy of association for protein complexes. *Protein Sci.* **1992**, *1* (1), 169–81.

(20) Ma, X. H.; Wang, C. X.; Li, C. H.; Chen, W. Z. A fast empirical approach to binding free energy calculations based on protein interface information. *Protein Eng.* **2002**, *15* (8), 677–81.

(21) Audie, J.; Scarlata, S. A novel empirical free energy function that explains and predicts protein-protein binding affinities. *Biophys Chem* **2007**, *129* (2−3), 198–211.

(22) Henriques, D. A.; Ladbury, J. E.; Jackson, R. M. Comparison of binding energies of SrcSH2-phosphotyrosyl peptides with structure-based prediction using surface area based empirical parameterization. *Protein Sci.* **2000**, *9* (10), 1975–85.

(23) Brooijmans, N.; Sharp, K. A.; Kuntz, I. D. Stability of macromolecular complexes. *Proteins* **2002**, *48* (4), 645–53.

(24) Camacho, C. J.; Vajda, S. Protein docking along smooth association pathways. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (19), 10636–41.

(25) Dell'Orco, D.; De Benedetti, P. G.; Fanelli, F. In silico screening of mutational effects on enzyme-proteic inhibitor affinity: a docking-based approach. *BMC Struct. Biol.* **2007**, *7*, 37.

(26) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.* **2005**, *48* (7), 2325–2335.

(27) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **2003**, *331* (1), 281–99.

(28) Hwang, H.; Pierce, B.; Mintseris, J.; Janin, J.; Weng, Z. Protein-protein docking benchmark version 3.0. *Proteins* **2008**, *73* (3), 705–9.

(29) Bordner, A. J.; Gorin, A. A. Protein docking using surface matching and supervised machine learning. *Proteins* **2007**, *68* (2), 488–502.

(30) Andrusier, N.; Nussinov, R.; Wolfson, H. J. FIREDOCK: fast interaction refinement in molecular docking. *Proteins* **2007**, *69* (1), 139–59.

(31) Liang, S.; Liu, S.; Zhang, C.; Zhou, Y. A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys. *Proteins* **2007**, *69* (2), 244–53.

(32) Lorenzen, S.; Zhang, Y. Identification of near-native structures by clustering protein docking conformations. *Proteins* **2007**, *68* (1), 187–94.

(33) Tovchigrechko, A.; Vakser, I. A. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* **2006**, *34* (Web Server issue), W310–4.

(34) Headd, J. J.; Ban, Y. E.; Brown, P.; Edelsbrunner, H.; Vaidya, M.; Rudolph, J. Protein-protein interfaces: properties, preferences, and projections. *J. Proteome Res.* **2007**, *6* (7), 2576–86.

(35) Li, C. H.; Ma, X. H.; Shen, L. Z.; Chang, S.; Chen, W. Z.; Wang, C. X. Complex-type-dependent scoring functions in protein-protein docking. *Biophys. Chem.* **2007**, *129* (1), 1–10.

(36) de Vries, S. J.; van Dijk, A. D.; Krzeminski, M.; van Dijk, M.; Thureau, A.; Hsu, V.; Wassenaar, T.; Bonvin, A. M. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **2007**, *69* (4), 726–33.

(37) Dominguez, C.; Boelens, R.; Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003**, *125* (7), 1731–7.

(38) May, A.; Zacharias, M. Protein-protein docking in CAPRI using ATTRACT to account for global and local flexibility. *Proteins* **2007**, *69* (4), 774–80.

(39) Zhou, H.; Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **2002**, *11* (11), 2714–26.

(40) Camacho, C. J.; Zhang, C. FASTCONTACT: rapid estimate of contact and binding free energies. *Bioinformatics* **2005**, *21* (10), 2534–6.

(41) Mashiach, E.; Schneidman-Duhovny, D.; Andrusier, N.; Nussinov, R.; Wolfson, H. J. FIREDOCK: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res.* **2008**, *36* (suppl_2), W229–232.

(42) Krissinel, E.; Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **2007**, *372* (3), 774–97.

(43) Cheng, T. M.; Blundell, T. L.; Fernandez-Recio, J. PYDOCK: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins* **2007**, *68* (2), 503–15.

(44) Pierce, B.; Weng, Z. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins* **2007**, *67* (4), 1078–86.

(45) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B.; Glen, R. C. Predicting protein-ligand binding affinities: a low scoring game. *Org. Biomol. Chem.* **2004**, *2* (22), 3267–73.

(46) So, S. S.; Karplus, M. A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13* (3), 243–58.

(47) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. 3rd, Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47* (12), 3032–47.

(48) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49* (20), 5912–31.

(49) Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.* **2008**, *29* (8), 1316–31.

(50) Eggers, C. T.; Wang, S. X.; Fletterick, R. J.; Craik, C. S. The role of ecotin dimerization in protease inhibition. *J. Mol. Biol.* **2001**, *308* (5), 975–91.

(51) Pornillos, O.; Alam, S. L.; Rich, R. L.; Myszka, D. G.; Davis, D. R.; Sundquist, W. I. Structure and functional interactions of the Tsg101 UEV domain. *EMBO J.* **2002**, *21* (10), 2397–406.

(52) Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins* **1999**, *35* (2), 133–52.

(53) Acampora, G.; Hermans, J., Jr. Reversible denaturation of sperm whale myoglobin. I. Dependence on temperature, pH, and composition. *J. Am. Chem. Soc.* **1967**, *89* (7), 1543–7.

(54) Zhou, H. X.; Rivas, G.; Minton, A. P. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* **2008**, *37*, 375–97.

(55) Schreiber, G.; Haran, G.; Zhou, H. X. Fundamental aspects of protein-protein association kinetics. *Chem. Rev.* **2009**, *109* (3), 839–60.

(56) Schreiber, G.; Fersht, A. R. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **1996**, *3* (5), 427–31.

(57) Camacho, C. J.; Kimura, S. R.; DeLisi, C.; Vajda, S. Kinetics of desolvation-mediated protein-protein binding. *Biophys. J.* **2000**, *78* (3), 1094–105.

(58) Gabdoulline, R. R.; Wade, R. C. Protein-protein association: investigation of factors influencing association rates by brownian dynamics simulations. *J. Mol. Biol.* **2001**, *306* (5), 1139–55.

(59) Li, S.; Xi, L.; Wang, C.; Li, J.; Lei, B.; Liu, H.; Yao, X. A novel method for protein-ligand binding affinity prediction and the related descriptors exploration. *J. Comput. Chem.* **2009**, *30* (6), 900–9.