PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Combining interface core and whole interface descriptors in postscan processing of protein-protein docking models

Noga Kowalsman[1] and Miriam Eisenstein[2]*

[1] Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel
[2] Department of Chemical Research Support, Weizmann Institute of Science, Rehovot 76100, Israel

ABSTRACT

Computational protein-protein docking scans commonly produce correct and nearly correct interaction models, but sometimes these models are ranked low. We present here postscan processing procedures that dramatically enhance the distinction between nearly correct and false predictions. The procedures employ propensity descriptors calculated for the interface core, an interface-core clusters count, solvation energy, and a geometric-electrostatic-hydrophobic complementarity score. The various descriptors rank high different selections of false models (shuffling effect), and therefore, are used as Boolean yes/no classifiers in "soft intersection" filters, which eliminate large proportions of false models. Furthermore, the standardized descriptors are used in new scoring functions that highlight nearly correct models (NCMs). All the tests are performed on unbound docking models produced with MolFit without use of external data. We find that the discrimination between nearly correct and false models by the various descriptors is class dependent; hence, our postscan processing is class specific. The filters reduce the number of putative models from 10,726, 12,517, and 11,054 to 758, 157, and 1218 for enzyme-inhibitor, antibody-antigen, and nonclassified systems. When combined with the new scoring functions, they improve the average rank of the highest ranking NCMs from 673 to 122. Application to 23 CAPRI targets demonstrates the effectiveness of the postscan procedures in cases where external information is used in the production of the putative models. Our new per-molecule residue propensity descriptors show that interacting interfaces are enriched with high propensity residues except for antigenic sites, which resemble more the noninteracting regions of protein surfaces.

Proteins 2009; 77:297–318.

Key words: unbound-unbound docking; MolFit; antigenic sites; voting classification; class specific scoring function; interface propensities; residue-residue propensities; solvation; CAPRI.

INTRODUCTION

The questions if and how proteins interact are of utmost interest in biological sciences because of the crucial roll that proteins play in almost every process in the living cell. Studies of the structures of protein-protein complexes and molecular recognition processes, therefore, can help to understand the cellular events that sustain life and lead to design and development of new drugs. Experimental methods such as yeast two-hybrid and mass spectroscopy are often used to test if given proteins interact; other experimental techniques such as site directed mutagenesis, crystallography, and NMR spectroscopy are commonly used to study how proteins interact. Computational docking routines, which predict the structures of complexes given the structures of the component molecules, also try to address this call. Ideally, docking procedures should be able to determine if proteins A and B interact but the current procedures struggle with the relatively simpler question "how do proteins A and B interact (given that they interact)?"[1]

Numerous docking algorithms have been presented in the literature (reviewed in Refs. 2–9). These algorithms search for the optimal interaction geometry by scanning and evaluating a large number of configurations and produce extensive lists of models. It is hoped that the lists include the correct model or one or more nearly correct models (NCMs), yet most of the putative models are false, some with high scores. Inclusion of biological and bioinformatics' data (*e.g.*, from mutagenesis experiments and sequence conservation analyses) either in the docking scan or in postscan filters is very beneficial for narrowing the ensemble of models and elevating the ranks of NCMs.[10,11] However, relevant external information is not always available and development of methods that enhance the discrimination between NCMs and false models based only on

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: The Kimmelman Center for Biomolecular Structure and Assembly.
*Correspondence to: Miriam Eisenstein, Department of Chemical Research Support, Weizmann Institute of Science, Rehovot 76100, Israel.
E-mail: miriam.eisenstein@weizmann.ac.il
Received 17 August 2008; Revised 13 March 2009; Accepted 14 March 2009
Published online 24 March 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22436

the properties of the interacting surfaces is therefore important.

Biologically relevant interfaces differ in many of their properties from the noninteracting portions of protein surfaces (*e.g.*,[12–14]) and from biologically irrelevant interfaces such as crystal contacts.[15] These properties (descriptors) can be used for discriminating between NCMs and false models. Some descriptors reflect the geometric properties of interfaces, such as the shape complementarity, the size of the interface, various packing and compactness measures, and general shape attributes (concave, convex, or flat).[12,14,16–19] A second group of descriptors represents the chemical character of the interface: its hydrophobicity,[12,16,20,21] the electrostatic complementarity of the interacting surfaces,[22–24] and the number of hydrogen bonds at the interface.[14] The physicochemical character of the interface also affects the interactions with the aqueous environment and it was shown that the solvation energy is higher for interfaces than for other parts of the surface.[25] A third group of descriptors is based on the statistical properties of interfaces, such as the relative number of polar and nonpolar groups,[2,14] residue propensities (RPs), which reflect the preferences of different amino acid types to be in the interface, and residue-residue propensities (RRPs) or pair wise potentials, which assess the tendency of two amino acids to interact with each other.[12,14,21,26–31]

Our groups' docking program, MolFit scans the rotation-translation space at regular intervals and evaluates each configuration with a geometric-electrostatic-hydrophobic (GEH) complementarity score. The GEH score estimates the shape complementarity and the size of the interface[17] together with the electrostatic[24] and hydrophobic complementarity.[20] We present in this article, a postscan processing procedure that employs additional interface descriptors: statistical RP and RRP descriptors and solvation energy (E_S). The RRP statistical potential reflects the physical forces that stabilize complexes. Hence, it includes specifics that also contribute to the MolFit scoring function, such as the tendency of hydrophobic residues to reside next to other hydrophobic residues and the preference of oppositely charged residues to form contacts with each other. However, RRPs also include information that is not accounted for by the GEH score, such as the favorable electrostatic interaction between aromatic rings and positive charges. Similarly, the solvation energy is related to the buried hydrophobic surface area but less related to the hydrophobic score of MolFit, which considers only complementary hydrophobic surfaces. Hence, the selected postscan descriptors generally differ from the descriptors that contribute to the GEH score and they have the potential to improve the distinction between NCMs and false models. Moreover, the postscan calculations presented here are not grid dependent and are not affected by grid-related scoring errors as seen for the GEH score.[32] Previously we showed that the GEH score is highly correlated with the interface size, and therefore, the distinction between NCMs and false models is more difficult when the predicted interface is small.[32] Therefore, normalized variants of the descriptors were tested and used, where necessary, to diminish the correlation with the interface size. We did not test the number of hydrogen bonds in the interface as a descriptor because this property is very sensitive to the exact relative positioning of the interacting molecules and to their conformation and is, therefore, less likely to be effective for identification and evaluation of NCMs in unbound docking (docking of molecules whose structures were determined in the uncomplexed form), in particular rigid-body unbound docking.

Propensity and solvation descriptors have been previously included in inscan scoring functions and in postscan filters (*e.g.*,[25,33–40]). Their implementation for evaluating unbound docking models is, however, error prone in particular for models produced by rigid-body docking. Thus, we anticipate that many of the correct residue-residue interactions are not formed even when the relative positioning of the docked molecules is exact, and spurious contacts are formed because of the unbound conformation. In this study, solvation and propensity values were calculated for the interface core (IC), which is structurally more stable than the rest of the interface,[41] and compared with descriptor values calculated for the whole interface. The IC consists of atoms that are totally buried in the interface and it is surrounded by a rim that includes atoms that are only partially buried upon complex formation.[42,43] The amino acid composition of the IC differs from the composition of the rim and of the noninteracting protein surface[42] and the residues in the IC are evolutionarily more conserved.[44] At the interaction level, the IC is the part that includes hot spot residues,[42,45,46] which contribute most to the interaction energy. We show here that IC descriptors, calculated with available RP and RRP tables, discriminate efficiently between NCMs and false models.

An important aspect of this study is the complexes' class partitioning. A common classification of complexes is based on their biological role, for example, enzyme-inhibitor (E-I), antibody-antigen (A-Ag), and electron transfer complexes. Several previous studies showed that some interface descriptors differ on the average for different classes of complexes.[12,21,28,31,47–51] Our results herein indicate that the discrimination between NCMs and false models by various IC and whole interface descriptors is class dependent. We took this information a step further and produced class specific filtering procedures and scoring functions. Only very recently, such scoring functions were proposed.[51–53]

Several discriminating IC and whole interface descriptors are combined here in new, class specific postscan filtering and re-evaluation procedures. The descriptors are used as Boolean yes/no classifiers in "soft intersection"

filters, which eliminate a large proportion of the false models reducing their average number from 10,695, 12,480, and 10,993 to 744, 153, and 1210 for E-I, A-Ag, and nonclassified systems ("other"), respectively, while keeping many NCMs. Thus, only one out of 63 systems loses all its NCMs in the filtering process. The elimination of models leads to improvement of the rank of the highest ranking NCM for 51 systems indicating that many of the discarded models are false positive; the average rank of these NCMs improves from 683 to 208 (for 62 systems). Re-evaluation of the filtered models by new scoring functions, which are linear combinations of standardized descriptor values, further improves the average rank of the highest-ranking NCMs to 122. Re-evaluation of all the GEH models with the new, class specific scoring functions improves dramatically the average rank of the highest-ranking NCMs from 673 to 198, for all 63 systems. Application of the new postscan processing procedures to 23 CAPRI (Critical Assessment of PRediction of Interactions)[54] targets leads to considerable improvement of the ranks of NCMs also in cases where external information was used in the prediction, and provides further validation of the effectiveness and robustness of these procedures. Hence, this study presents a major enhancement of our ability to identify correct and nearly correct docking models and rank them high. The robustness of our procedures suggests that they are likely to be effective for docking decoys produced by a variety of algorithms.

METHODS

Database of unbound-unbound docking models

We constructed a database of unbound-unbound docking models for a diverse benchmark of systems commonly used by the docking community.[55] The coordinates of each molecule were prepared as previously described for bound systems.[32] MolFit rotation-translation scans were executed employing the previously determined standard parameters (translation interval of 1.05 Å, rotation interval of 12°, saving 2 models per orientation[32]), producing three ensembles of docking models for each system: geometric,[17] geometric-electrostatic,[24] and geometric-hydrophobic.[20] The lists of models were intersected (GEH intersection) producing a final list that included only models that appeared in all three lists. These models were evaluated by a combined GEH score, composed of the geometric score + electrostatic score + hydrophobic score. All the GEH models for each system were compared with the experimental structure of the native complex to detect NCMs. We used the CAPRI definition of NCMs,[56,57] which distinguishes between three levels of NCMs' accuracy (***, **, and *) according to the ligand root mean square deviation (RMSD), interface RMSD, and fraction of native contacts (*Lrmsd, Irmsd,* and *Fnat*, respectively). Throughout this article, we refer to the rank of the highest ranking NCM for a given system as the "rank of the best NCM" except where otherwise indicated.

Detection of the interface core using accessible surface area calculations

The interface core (IC) atoms of each docking model were determined by comparing their water accessible surface area (ASA) in the monomer and in the model. The ASA of the component molecules and of the docking models were calculated with the widely used program NACCESS.[58] NACCESS slices each atom according to the value of the parameter *Z-slice* and calculates for each slice the length of the exposed arc. The atoms are sliced along the *z*-axis and the calculated ASA is, therefore, dependent on the orientation of the molecule relative to the coordinates system. Our tests indicated that for the default setting of NACCESS (*Z-slice* = 0.05 Å) the per atom ASA values calculated for different orientations of the same molecule deviate by up to 2.0 Å^2 (0.6 Å^2 on average). Deviations of such magnitude may introduce errors in the selection of IC atom. Therefore, we set *Z-slice* to 0.01 Å, thus, reducing the breadth of deviations to 0.3 Å^2 (0.1 Å^2 on the average). Also, we included in the ASA calculations heterogroups that are part of the docked proteins structures.

The IC was defined by Chakrabarti and Janin[42] as the ensemble of atoms which become totally buried in the interface upon complex formation. This definition may be too strict for the evaluation of rigid body docking models, particularly in unbound-unbound docking. We therefore introduced a maximum exposure parameter, EX_{max}, and in the first step of the IC detection procedure we selected interface atoms (atoms that change their ASA upon complex formation) with ASA $\leq EX_{max}$. Next, these atoms were clustered, requiring that each atom in the cluster resides within a distance D_c from any other atom in that cluster. We surmised that the clustering process would retain IC atoms and residues that are part of continuous binding patches and reduce the number of false IC assignments. Such false assignments are likely to occur for rigid body docking models because the conformations of the interacting molecules are not modified. The final IC consisted of all the clusters that included at least N_c atoms. Notably, the IC detection was executed separately for the interface atoms of each molecule in the complex and for both molecules. IC residues were defined as residues that have at least NA interface core atoms. Hence, the implementation of the IC detection procedure required calibration of five free parameters: EX_{max}, N_c, D_c, and NA that are defined above and a fifth parameter (BB) that controls if all the IC residues con-

tribute to the propensity calculations or only residues that have at least one side-chain atom buried in the IC.

Calibrating the free parameters in the interface core propensities calculations

The calibration of the free parameters in the IC properties evaluation program (ICp) consisted of two independent rounds using a different set of 30 systems in each round. The two sets had similar composition of complexes in terms of their biological role, a wide range of interface sizes and a similar number of higher accuracy NCMs (** and ***), and acceptable NCMs (*). The calibration benchmark consisted of 110 models per system, including 1–50 NCMs, 50–99 highest-ranking false models, and 10 false models with low MolFit ranks. These models were divided into four bins as follows: bin1 included the more accurate NCMs, bin2 contained acceptable NCMs, the high ranking false models were inserted into bin3, and the low ranking false models were put in bin4. For each setting of the ICp parameters propensity values were calculated for each model and the average for each bin was determined; then the differences between the averages (bin1-bin3, bin2-bin3, and bin4-bin3) were calculated for each system and finally the average differences for the 30 systems in the calibration set were calculated.

Different settings of the ICp parameters were tested except for the parameter NA, which was set to 1 based on several preliminary tests. EX_{max} was set to 0.2 $Å^2$ in the calibration calculations, which corresponds to including only totally buried atoms in the IC (within twice the average ASA calculation error). The optimal parameters were those that conferred the largest differences between bin3 and bin2 or bin1. More emphasis was given to the difference between bins 2 and 3, which represents the level of discrimination between low accuracy NCMs and false positive models. The ICp parameters were calibrated separately for each of the propensity descriptors described below, except that the optimal parameters for IC-RP_{AB} were also used in the calculation of the new per-molecule descriptors IC-RP_A and IC-RP_B because IC-RP_{AB} includes a larger number of IC residues, improving the statistics of the calibration.

The interface core and whole interface descriptors tested in this study

The different IC and whole interface descriptors tested in this study are listed in Table I. Three IC residue propensities (IC-RP) were calculated for each model, IC-RP_A, IC-RP_B, and IC-RP_{AB}, by summing the propensity values of residues included in the IC of molecule A, molecule B, and both molecules, respectively. We used the RP values determined for the IC by Chakrabarti and Janin.[42] Positive values indicate that the tested interface comprises residues with high preference to be in the IC.

We tested three tables of RRP values, referred to below as RRP_1[29], RRP_2[27], and RRP_3 (log of the preference fractions of Anashkina *et al.*[30]). These tables differ in the definition of the residue-residue contacts and the expected residue frequencies. Thus, Lu *et al.*[29] (RRP_1) consider two interface residues to be in contact if any nonhydrogen atom of residue i is within 5 Å of a nonhydrogen atom of residue j in the other molecule, and the expected frequency is based on the fraction of residues of type i or j out of the total number of surface residues. Glaser *et al.*[27] (RRP_2) count i, j contacts with $C\beta_i$-$C\beta_j$ distances < 6 Å, and the expected frequency is the fraction of residues of type i out of the total number of interface residues. Anashkina *et al.*[30] (RRP_3) count only physical contacts (using Voronoï tessellation) and the expected frequency is the fraction of residues i and j out of the total number of contacting residues.

The three RRP tables were originally derived for whole interfaces. They are used in this study for the calculation of whole interface and IC residue-residue propensities (WI-RRP and IC-RRP, respectively) because we did not find RRP tables derived for the IC. IC-RRP values were calculated by summing the RRPs for all the interacting pairs of IC residues. Residues i and j were considered an

Table I
The Different Descriptors Tested in this Study

Descriptor	Abbreviation	Description
Interface core residue propensities	IC-RP_{AB}	Residue propensity of the interface core according to Ref. 42
	IC-RP_A	Residue propensity of the interface core of molecule A
	IC-RP_B	Residue propensity of the interface core of molecule B
Interface core residue-residue propensities	IC-RRP_1	Residue-residue propensity of the interface core according to Ref. 29
	IC-RRP_2	Residue-residue propensity of the interface core according to Ref. 27
	IC-RRP_3	Residue-residue propensity of the interface core according to Ref. 30
Interface core desolvation energy	IC-E_s	The desolvation energy of the interface core calculated according to Ref. 59
Interface core atoms	IC_{NA}	The number of atoms in the interface core
Interface core residues	IC_{NR}	The number of residues in the interface core
Interface core clusters	IC_{NC}	The number of atom clusters in the interface core
Whole interface residue propensities	WI-RRP_1	Residue-residue propensity of the whole interface according to Ref. 29
	WI-RRP_2	Residue-residue propensity of the whole interface according to Ref. 27
Whole interface desolvation energy	WI-E_s	The desolvation energy of the whole interface calculated according to Ref. 59

interacting pair if at least one nonhydrogen atom from residue i in molecule A resided within a distance D_{RR} from a nonhydrogen atom of residue j in molecule B. D_{RR} was calibrated together with the other ICp parameters. WI-RRP_1 and WI-RRP_2 were calculated for all the interface residues, employing the same definition of interacting pairs as in the derivation of these RRP tables. The three original RRP tables were scaled differently; we rescaled the IC-RRP and WI-RRP values for convenience, bringing all the propensities to the same range of magnitudes. Also, the signs of the RRP values were reversed hence, in this article, positive IC-RRP and WI-RRP values denote favorable contacts.

The solvation energy of the IC (IC-E_S) was calculated as the sum over all the IC atoms (nonclustered) of the products of atomic solvation parameters[59] and atomic surface areas buried upon complex formation. Whole interface solvation energies (WI-E_S) were calculated in the same manner for all the interface atoms. More positive solvation energies indicated larger contribution towards the stability of the complex.

Normalized variants of each of the propensity and solvation descriptors described earlier were also tested. These were expected to be less dependent on the interface size. IC-RP values were divided by the number of IC residues; IC-RRP and WI-RRP values were divided by the number of interacting residue pairs and the IC-E_S and WI-E_S were divided by the combined ASA of the desolvated atoms. In addition to the propensity descriptors and the solvation energy, we also tested the number of IC atoms (IC_{NA}), residues (IC_{NR}), and atom clusters (IC_{NC}) for their ability to distinguish between NCMs and false models.

Evaluation of the performance of the interface descriptors

Descriptors can be used as Boolean yes/no classifiers by specifying a threshold value that segregates the models into positive (potentially correct) and negative (potentially false) models. Several measures are commonly used to evaluate the performance of a classifier for a specific threshold, such as the sensitivity (Sn) and the specificity (Sp). Sn and Sp are defined as Sn = tp/(tp + fn) and Sp = tn/(tn + fp), where tp (true positive) is the number of NCMs above the threshold, tn (true negative) is the number of false models below the threshold, fn (false negative) is the number of NCMs below the threshold and fp (false positive) is the number of false models above the threshold. Hence, Sn is the fraction of NCMs above the threshold out of the total number of NCMs (true positive rate) and Sp is the fraction of false models below the threshold out of all the false models (true negative rate). The performance of a descriptor is evaluated by plotting a receiver operating characteristic (ROC) curve, which presents Sn *versus* 1-Sp (*i.e.*, true positive rate *vs.* false positive rate) for different threshold values. The curve for a nondiscriminative (random) descriptor is a diagonal line with a slope of 1, and the area under this curve (AUC) is 0.5. The curve is above the diagonal for discriminative classifiers and the AUC increases as the discrimination ability of the descriptor improves.[60]

Sn and Sp for selected thresholds were calculated for each of the 63 systems for which MolFit produced NCMs, and averaged. Simple averaging may be incorrect when the distributions of descriptor values are system dependent and the thresholds have different statistical significance for each system. Therefore, the distributions of descriptor values were matched by shifting the mean of each distribution to 0.0 ($Z1$) or by further dividing $Z1$ by the standard deviation (STD) of the distribution. These are referred to below as the shifted distributions $Z1_i = X_i - < X >$, where X_i is the descriptor value for model i and $< X >$ is the average of the distribution, and the standardized distributions $Z2_i = Z1_i/\mathrm{STD}$.

Intersection and "soft intersection" filters

The intersection filter tests lists of models ranked by different descriptors and retains models that appear in all the lists. The filter can be applied to models with descriptor values above selected thresholds or to a preset number of top ranking models. We introduce here a "soft intersection" filter (a voting classification procedure), which retains models with descriptor values above the selected thresholds of any N out of M descriptors. When $N < M$ the filter is less demanding and the test is termed "soft intersection". N, M, and the optimal thresholds were determined for classes of complexes divided by their biological role. We considered two criteria: First, the number of systems that lost all their NCMs was smallest and the fraction of NCMs retained by the filter was high. Second, the NCMs enrichment ratio, calculated as the ratio between the percent of NCMs and percent of false models retained by the filter was highest. We consistently preferred parameters that retained more NCMs (at the cost of also retaining more false models) because a softer filter is more likely to be adequate for a wider range of systems than in the current database.

New scoring functions for evaluating docking models

The standardized descriptor values, $Z2$, were used to construct class specific scoring functions in the form of linear combinations. The coefficients of the linear combinations were optimized by extensive step-wise searches. We first conducted comprehensive coarse searches, with the coefficients set to combinations of 0., 0.25, 0.5, 0.75, and 1. These were followed by fine searches in which the coefficients were incremented in steps of 0.05. The optimization of the coefficients was done against filtered lists

of models in which the ratio between the number of NCMs and false models is much higher than in the MolFit GEH lists. For each setting of the coefficients we calculated new scores and redetermined the rank of the best NCM for each system. The average of these ranks for all the systems was used to assess the quality of the scoring function.

Cross validation of the optimized "soft intersection" filters and scoring functions

Five-fold cross validation tests were used to establish the statistical validity of the optimized post scan procedures. Thus, the systems in each class were divided randomly into five groups and the parameters optimization procedures were repeated five times using different selections of four out of the five groups. The optimal parameters in each case were applied to the left out group. The first test validated the optimized "soft intersection" filters using the average NCMs enrichment ratio and the average rank of the best NCMs as performance measures. The second test validated the scoring function coefficients using the average rank of the best NCMs and the average AUC as performance measures.

RESULTS

NCMs in the unbound-unbound docking database

Unbound-unbound docking scans were executed with our docking program MolFit for 69 systems from the benchmark of Mintseris *et al.*,[55] and all the docking models were assessed using the CAPRI assessment criteria.[56,57] To estimate the prediction difficulty of a system (Table II and Supporting Information S1), we produced a "superposition model" by superposing each of the unbound molecules on its counterpart in the complex, and calculated the *Fnat* and *Irmsd*. The *Lrmsd* for the superposition model is small, in the range 0.4–3.9 Å for the 69 systems used in this study. Hence, although the *Lrmsd* is an essential measure for assessing docking models it is not very useful for assessing the prediction difficulty of a system. In the most favorable case rigid body docking can form high accuracy (***) NCMs for only 33 systems. This result together with the expected positioning error for a grid interval of 1.05 Å, suggest that high accuracy models can be obtained in only few cases. The *Fnat* values for the superposition models range between 0.19 and 0.90 (0.64 $\pm$ 0.15 on the average); small *Fnat* indicates that even at the correct relative positioning of the molecules the interface properties calculations are likely to carry considerable errors.

Table II
MolFit Prediction Results for 69 Unbound-Unbound Docking Systems

	All systems	E-I	A-Ag	``other''
Number of systems	69	23	10	36
Prediction difficulty[a]	33, 24, 12	14, 6, 3	6, 4, 0	13, 14, 9
Number of cases with NCMs	63	23	9	31
Average number of NCMs	27	35	37	18
Average rank of the best NCMs	673	323	180	1077
Distribution of the most accurate NCMs[b]	7, 36, 20	5, 14, 4	0, 6, 3	2, 16, 13

[a]The number of superposition models with accuracy ***, ** and * is listed.
[b]The number of systems for which the most accurate model is ***, ** or *.

The GEH intersection was previously shown to be very efficient in unbound docking as it eliminates a large fraction of false models and improves significantly the ranking of NCMs.[20] The GEH lists in this study consist of 5514–14,297 models per system (11,178 models on average out of 17,520 models per scan) and include NCMs for 63 of the 69 systems. MolFit failed to form NCMs for two cases with (***) superposition models, 1sbb and 1ghq. These systems have particularly small interfaces (1064 Å^2 and 1245 Å^2, respectively), which were previously shown to be hard to detect.[32,61] NCMs were not formed also for 1bgx, which has a (**) superposition model and a large, potentially easy to detect, interface (5814 Å^2). In this case the unbound antibody includes four additional residues at the N-terminus of the heavy chain, which clash with the antigen and prevent formation of NCMs. The three other systems for which NCMs were not formed are difficult systems with low accuracy (*) superposition models (1atn, 1ibr, and 1fak).

The MolFit scans in this study employ sparse sampling of the rotation-translation space, and therefore, the number of NCMs is usually small, 27 NCMs on average per system. As expected, *** models are obtained in only seven cases and for most systems only ** and * models are formed (Table II). The best NCMs for the 63 systems show a wide range of ranks (Table S1) but in 40 cases the ranks are below 250 (18/23 E-I, 6/9 A-Ag, and 16/31 "other" systems). The average ranks of the best NCMs are better for E-I and A-Ag systems than for "other" (Table II), in line with the generally lower accuracy of the superposition models for the latter class.

The distributions of descriptor values for NCMs and false models overlap

The ICp free parameters were calibrated as described in the Methods section and the results of the calibration are summarized in Supporting Information S2. Normalized and non-normalized descriptor values were calculated for all the GEH docking models of 63 unbound-unbound systems for which at least one NCM was formed, employing the calibrated ICp parameters listed in Table S2. The distributions of descriptor values for the NCMs and false models of several systems were examined manually to obtain initial assessment of their discrimination ability. Examples of discriminative and nondiscrimi-

native cases are shown in Figure 1 for one E-I and one "other" system. The maxima of the distributions for false models and for NCMs are significantly displaced for IC-RP_{AB} and IC-RRP_1, indicative of their discrimination power. The results for the other descriptors are less conclusive and system dependent, but not class specific. Importantly, the distributions for the NCMs and for false models overlap to some extent in every case hence none of the descriptors provides unambiguous distinction between NCMs and false models on its own.

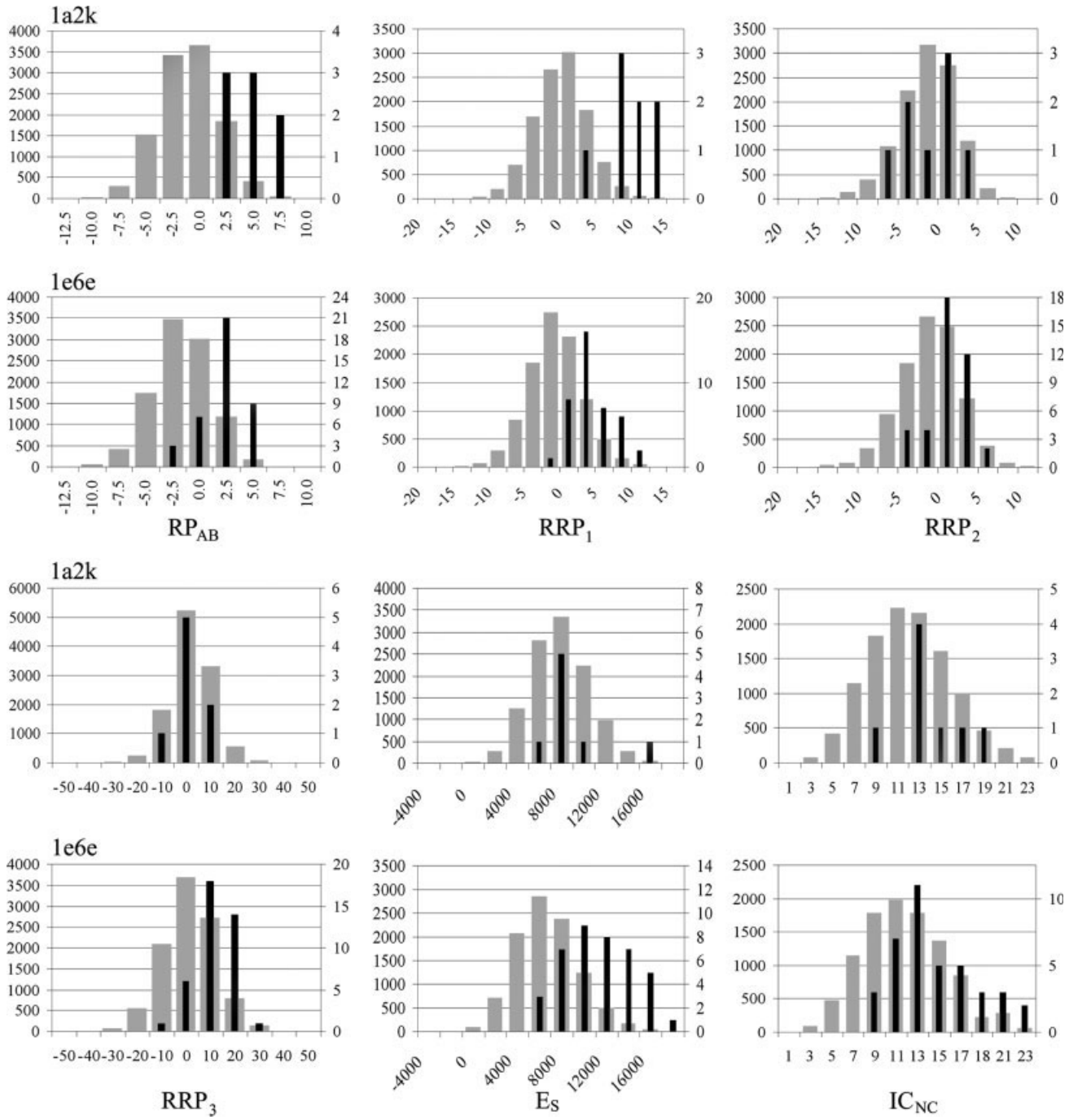

Figure 1

The distribution of descriptor values for two systems, 1a2k ("other" class) and 1e6e (E-I class). The gray columns depict the number of false models with descriptor values in the given range and the black columns indicate the number of NCMs in the same descriptor ranges. Note the different scales for the number of NCMs (on the right) and of false models (on the left).

Further examination of the distributions of descriptor values indicates that the averages of the bell shaped distribution are system dependent and for some descriptors the range of the averages is wide. For example, the average non-normalized IC-RP_{AB} (for EX_{max} = 5 Å^2) range between −7.83 and 0.97. Interestingly, the range of the STDs for the same distributions is narrow (2.86 ± 0.34), indicating that the shapes of the distributions are similar for the 63 systems but they are shifted along the horizontal (descriptor value) axis. Such shifts are observed to some extent for all the descriptors (data not shown) and they must be taken into consideration in the selection of threshold values that render a model potentially correct or potentially false (see below). The similar shape of the distributions for different systems reflects the similar composition of the surface residues in these water soluble molecules and the preferences of the scoring function used in the MolFit scans. The different average descriptor values may be related to differences in the size and shape of the molecules.

Can the interface core descriptors distinguish between NCMs and false models?

The ability of a given descriptor to discriminate between NCMs and false models is assessed by plotting average ROC curves. Thus, Sn and 1-Sp for selected descriptor thresholds are calculated for all the systems, then average Sn and 1-Sp are determined for each threshold together with the standard deviations (SDs). Interestingly, the SDs for Sn are significantly higher than the SDs for 1-Sp (calculated for either *Z*1 or *Z*2 or the actual descriptor values) reflecting the greater diversity in the number of NCMs used in the calculation of Sn compared with the number of false models considered in the 1-Sp calculations. Also, the SDs are consistently smaller for shifted and standardized distributions (*Z*1 and *Z*2) than for the actual descriptor values, indicating that *Z*1 and *Z*2 vary less between systems. Moreover, the ROC curves for *Z*1 and *Z*2 are practically identical, further supporting the abovementioned observation that the distributions of a given descriptor values are very similar for all the systems except for a horizontal shift. The *Z*1 values are therefore used in the following ROC curve analyses, ascertaining that each threshold has the same statistical significance for all 63 systems.

The ICp parameter EX_{max} determines which atoms can be included in the IC. Optimal EX_{max} values were determined by comparing the averaged ROC curves (Fig. 2) and their SDs (Supporting Information S3). We found that an increase in EX_{max} from 0.2 Å^2 to about 5 Å^2 improves the AUC for most IC descriptors. Presumably, the larger EX_{max} confers adequate tolerance in the detection of IC atoms for models produced by unbound rigid-body docking. Additional increase of EX_{max}, up to 15 Å^2, marginally improves the AUC for IC-RRP_1. IC-RRP_2 is the least discriminative IC-RRP descriptor for any EX_{max} value (Fig. 2 and Table S3). WI-RRP_2 values calculated for 20 systems that include E-I, A-Ag, and "other" show that this descriptor is nondiscriminative (AUC = 0.51). The calculations of IC-RRP_2 and WI-RRP_2 were therefore discontinued.

In contrast to the propensity descriptors, the solvation energy is clearly more discriminative when calculated for the whole interface (WI-E_S) than for the IC (Fig. 2 and Table S3); it is however a rather weak descriptor with AUC = 0.67 for WI-E_S. Normalization has a very small effect on the ROC curves for all the descriptors and for all EX_{max} values (Table S3).

The detection of the IC produced three additional putative descriptors: IC_{NA}, IC_{NR}, and IC_{NC}. These statistical properties were calculated for EX_{max} = 5 Å^2. The descriptors IC_{NA} and IC_{NR} are weak discriminators, with AUC = 0.56 and 0.58, respectively; IC_{NC} has more discriminative power, with AUC = 0.63 (Fig. 3).

Selecting the best descriptors

None of the descriptors can unambiguously distinguish between NCMs and false models. Similarly, previous studies showed that single interface descriptors cannot fully distinguish between biologically relevant interfaces and crystal contacts.[2,15,62–64] We therefore considered the possibility of combining descriptors, provided that they are not too strongly correlated. The Pearson correlation coefficient, *r*, was calculated for all pairs of descriptors for 20 systems, which included E-I, A-Ag, and "other", and averaged (Supporting Information S4). Most pairs of descriptors are not correlated and the absolute *r* values are below 0.3. Moreover, in most cases very similar *r* are obtained for pairs of normalized descriptors and corresponding pairs of non-normalized descriptors. The exception is WI-E_S, for which normalization lowers the correlation with the MolFit GEH score, reducing the average *r* from 0.31 to 0.10. The GEH score is dominated by the contribution of the geometric complementarity term and was shown to be correlated with the interface size.[32] Hence, the correlation between the non-normalized WI-E_S and the GEH score is a result of the correlation of both descriptors with the interface size. The descriptors IC_{NA} and IC_{NR} are correlated with the GEH score (*r* = 0.49 for both) and weakly correlated with the non-normalized WI-E_S but not with the normalized WI-E_S, indicating that they too are interface size dependent. Interestingly, the number of IC clusters, IC_{NC}, is much less correlated with the GEH score (*r* = 0.23) than IC_{NA}, indicating that the average size of the IC clusters generally increases with the size of the interface. IC-RP_{AB} and IC-RRP_1 are signifi-

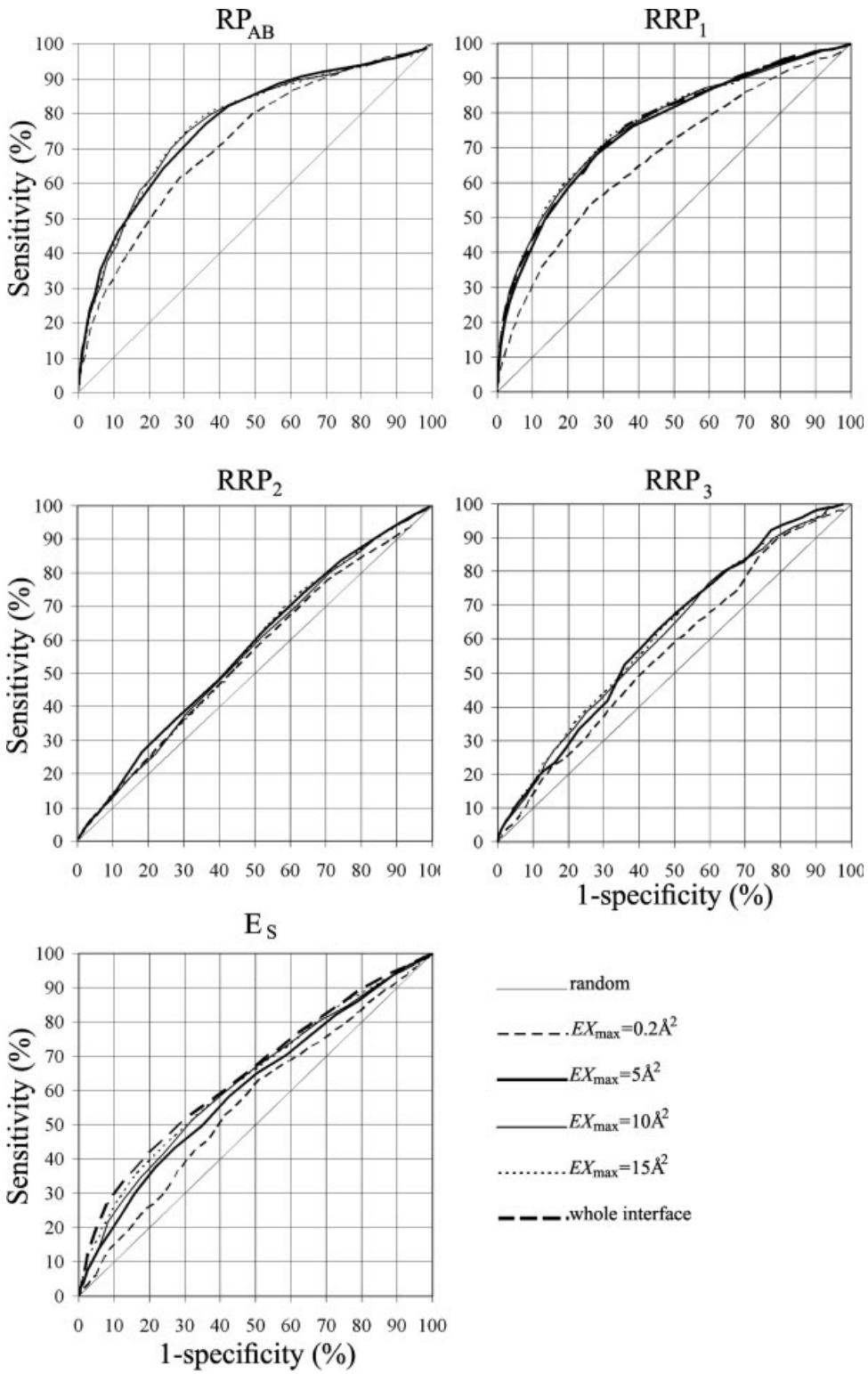

Figure 2

ROC curves for interface core and whole interface propensity and solvation descriptors. Interface core descriptors calculated with several EX_{max} values are compared.

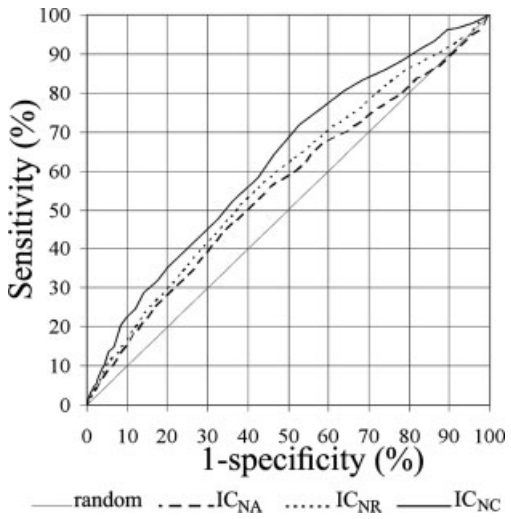

Figure 3

ROC curves for the interface core statistics descriptors.

cantly correlated ($r = 0.72$) probably because the expected residue frequency for the RRP_1 table is based on the fraction of residues of type i on the surface and not only in the interface;[29] hence this RRP table is contaminated with RP information.

Based on the ROC curves, the comparison of normalized and non-normalized descriptors and the correlation analyses we chose the following descriptors for further testing: (1) The non-normalized IC-RP_{AB} calculated with $EX_{max} = 5$ Å^2 and the per-molecule descriptors IC-RP_A and IC-RP_B; the latter are relevant only for the E-I and A-Ag classes where molecules A (enzyme or antibody) and B (inhibitor or antigen) can be unambiguously assigned. (2) The non-normalized IC-RRP_1 calculated with $EX_{max} = 15$ Å^2; the larger EX_{max} is in line with the fact that the RRP_1 table was derived for whole interfaces. Although IC-RRP_1 and IC-RP_{AB} are significantly correlated, each of them may carry independent information and both descriptors were therefore retained. (3) The non-normalized IC-RRP_3, calculated with $EX_{max} = 10$ Å^2; this is a weak discriminator however it is not correlated with IC-RP_{AB} or IC-RRP_1 suggesting that it may be useful in combinations of several descriptors. (4) The normalized WI-E_S. (5) The IC statistical descriptor IC_{NC}, which is more discriminative than IC_{NA} and IC_{NR} and is not significantly correlated with the GEH score. (6) The MolFit GEH score. Standardized scores (GEH-Z) were calculated as GEH-Z = (GEH score $- \mu)/\sigma$, where μ and σ are the mean and SD of the distribution of scores for a given system, and are obtained by fitting an extreme value distribution function to the observed distribution.[32,65] The AUC for GEH-Z (0.72) indicates that it is a discriminative descriptor.

Different descriptors are effective for different classes of complexes

The somewhat different interface characteristics for different classes of complexes, such as obligatory and nonobligatory oligomers or E-I and A-Ag systems, have been previously discussed.[12,14,21,31,48,49,66] We therefore tested the behavior of the descriptors described here for different classes of complexes divided according to their biological role, the size of the native interface or its chemical character.

Classification by the biological role of the complex

The 63 unbound-unbound systems considered here include 23 E-I and 9 A-Ag systems. The remaining 31 "other" systems include too few cases with similar biological role to be grouped and produce statistically sound results and conclusions. ROC curves for the E-I and A-Ag classes are shown in Figure 4 and the AUC are listed in Table III. IC-RP_{AB}, IC-RRP_1, IC-RRP_3, and GEH-Z are discriminative for both classes, whereas WI-E_S and IC_{NC} are discriminative only for the E-I class. The per-molecule descriptors, IC-RP_A and IC-RP_B, clearly differentiate between the E-I and A-Ag classes. Both are discriminative descriptors for E-I systems, with AUC of 0.79 and 0.70, respectively. These results probably reflect the evolutionary processes that the enzyme and the inhibitor underwent to ensure better activity and specificity. Utterly different behavior is observed for the A-Ag class, where IC-RP_A, calculated for the antibodies, is very discriminative (AUC = 0.84), whereas IC-RP_B, calculated for the antigen, is nondiscriminative. These results are also biologically sensible; the antibody undergoes an evolution-like process to optimize its interactions with the antigen. In contrast, antigens do not seek to be recognized by antibodies.

Notably, the selection of residues in antigenic sites is not random. The shape of the ROC curve for IC-RP_B and its AUC (0.42) suggest that antigenic sites avoid residues that are preferred in other interfaces. The antigenic site propensities derived by Haste *et al.*[67] support this notion. Hence, some of the more common residues in antigenic sites, Ala, Pro, Lys, and Gln, have particularly low IC-RP; in fact, Lys is the least abundant IC residue.[42] Correspondingly, residues that are preferred in the IC such as Cys and Ile are rarely found in antigenic sites. We found that the antigenic site propensities do not discriminate between NCMs and false models in A-Ag systems producing an approximately diagonal ROC curve with AUC of 0.52. This weak signal is further supported by our analysis of the interfaces in the nine A-Ag systems, which indicates that only 55% of the contacts (62 of the 112 contacts below 3.5 Å) are formed by residues with high antigenic site propensities. The available residue propensity data for protein surfaces,[68] interfaces,[14] ICs,[42] and antigenic sites[67] have different scales

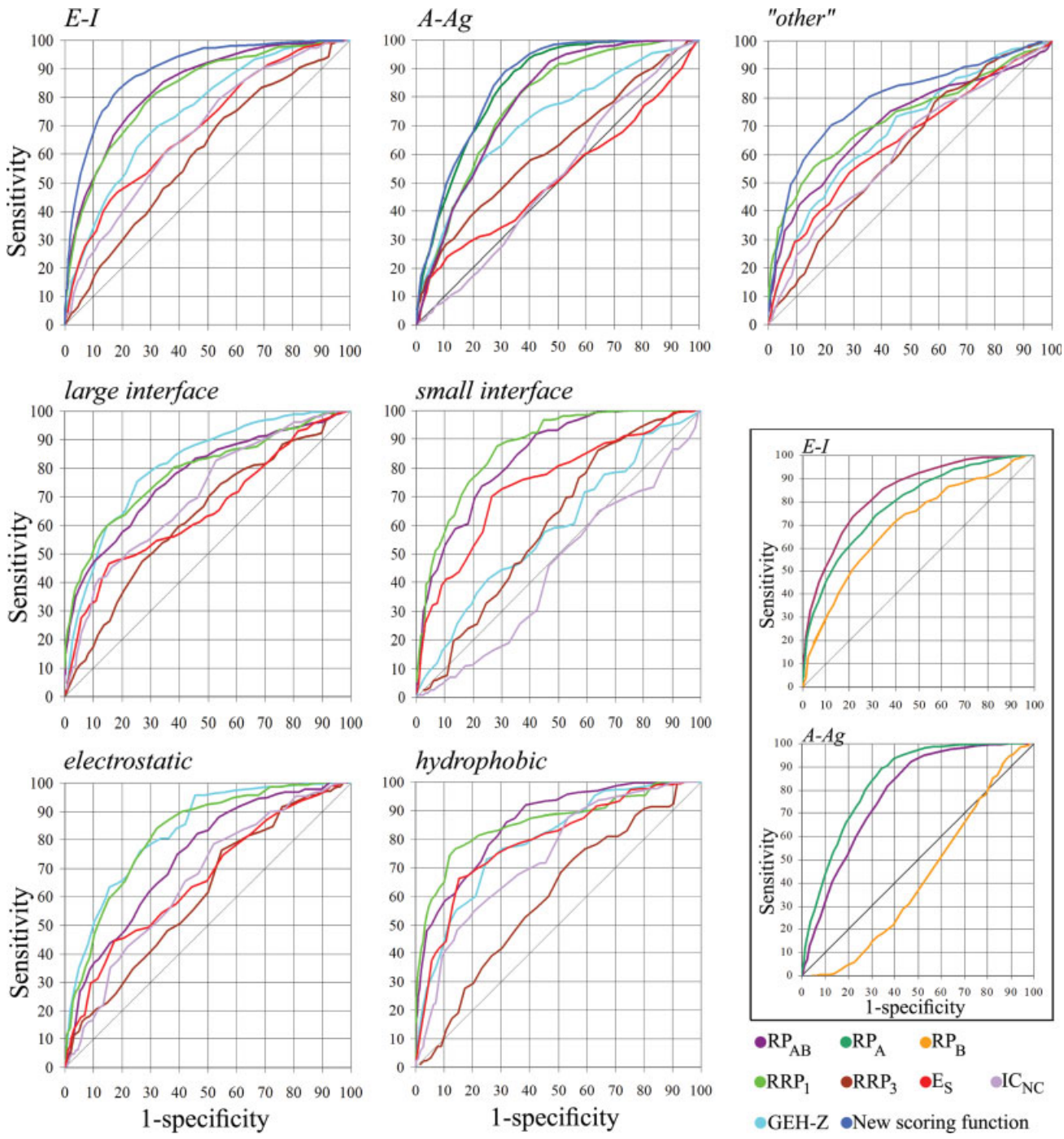

Figure 4

ROC curves for classes of systems divided by their biological role or by the size or chemical character of the native interface. The insert on the right compares the performance of the per-molecule descriptors IC-RP_A and IC-RP_B to IC-RP_{AB} for the E-I and A-Ag classes.

precluding detailed comparison; but the signs of the propensities are scale independent and we note that the signs of surface propensities and antigenic site propensities are the same for 12 amino acids, 10 of which have interface and IC propensities with opposite signs. Hence, the composition of antigenic sites is more similar to that of the noninteracting surface portions than to interfaces or ICs, making them inadequate for distinguishing between NCMs and false models. The analysis of 9 A-Ag interfaces also shows that 16 antigen Asn residues are in contact with the antibody in 8 of the 9 interfaces and 15 antigens Lys residues are found in 7 interfaces. General conclu-

Table III
AUC Values for the ROC Curves for Different Descriptors for All 63 Systems and for Classes of Systems

Descriptor	All systems	Biological role			Interface size		Chemical character	
		E-I	A-Ag	``other''	small	large	elec.	hydr.
GEH-Z	0.72	0.75	0.72	0.69	0.59	0.81	0.83	0.79
IC-RP_{AB}	0.77	0.84	0.78	0.71	0.84	0.77	0.74	0.86
IC-RRP_1	0.77	0.82	0.78	0.73	0.87	0.79	0.82	0.86
IC-RRP_3	0.61	0.60	0.62	0.62	0.61	0.63	0.61	0.60
WI-E_S	0.65	0.70	0.52	0.65	0.74	0.65	0.66	0.79
IC_{NC}	0.63	0.68	0.52	0.62	0.45	0.70	0.66	0.74
IC-RP_A	–	0.79	0.84	–	–	–	–	–
IC-RP_B	–	0.70	0.42	–	–	–	–	–
New scoring function	0.84	0.90	0.85	0.79				

sions cannot be drawn based on only nine A-Ag systems yet the pronounced abundance of Asn and Lys on the antigen side of the interface can perhaps be used for identifying NCMs in A-Ag docking.

WI-E_S is another descriptor that differentiates between the E-I and A-Ag classes (AUC values of 0.70 and 0.52, respectively). Our results differ from the results of Burgoyne *et al.* who calculated the desolvation energy of the clefts in protein-protein interfaces. They found that although the desolvation energy is significantly less discriminative for A-Ag and "other" than for E-I systems it is nevertheless a strong descriptor for antibodies and a reasonable descriptor for antigens.[66] This difference is probably a result of the way the E_S was calculated - normalized whole interface E_S in this study as opposed to E_S of only the clefts in the other study. IC_{NC} too is discriminative for the E-I systems but not for A-Ag systems. Perhaps this is related to the shapes of the interfaces; in E-I complexes the inhibitor binds in a cavity on the surface of the enzyme, whereas the interfaces of A-Ag complexes with protein antigens are generally flat.

The class of "other" is not uniform in terms of the biological role of the molecules. It however consists of about half of the systems in the unbound-unbound database and we therefore analyzed the performance of the descriptors for this class (Fig. 4 and Table III). GEH-Z is less discriminative for "other" than for the E-I and A-Ag classes, in line with the lower MolFit ranking of the NCMs (Table II). A similar trend is observed for IC-RP_{AB} and IC-RRP_1 yet in general, the shapes of the ROC curves for the "other" class are more similar to the corresponding curves for the E-I class than for the A-Ag class.

The results above strongly suggest that different descriptors should be used for filtering and re-evaluation of models for the different classes. For example, there is no point in including E_S in the evaluation of A-Ag systems and it will be better to use IC-RP_A instead of IC-RP_{AB} for this class.

Classification by the size or the chemical nature of the native interface

We compared the behavior of the descriptors for systems with particularly small or particularly large interfaces, and systems with large electrostatic or hydrophobic complementarity. These analyses cannot be used in predictive docking because we do not know in advance the properties of the interface being predicted; their purpose is to detect trends in the behavior of the descriptors.

Analyses of the interfaces in a large ensemble of heterocomplexes indicated that the average interface size is 1600 ± 400 Å^2.[14,42] We compare the ROC curves for a class of complexes with small interfaces (10 systems with interface area <1350 Å^2; 2 E-I and 8 "other") and a class of complexes with large interfaces (21 systems with interface area >2000 Å^2; 7 E-I and 14 "other"). A-Ag systems were excluded from this analysis because of their different behavior described above. As expected, GEH-Z is a weak discriminator for the small interfaces class (AUC = 0.59, Fig. 4 and Table III) and a good descriptor for the large interfaces class (AUC = 0.81). In contrast, IC-RP_{AB}, IC-RRP_1, and WI-E_S are significantly more discriminative for small interfaces than for large interfaces. This result is not related to the composition of the small interfaces class, which consists mostly of "other" systems for which these descriptors are generally less discriminative. Possibly, native small interfaces are enriched with preferred IC residues and residue contacts that stabilize these complexes. The preferred IC residues are mostly hydrophobic explaining the better discrimination ability of WI-E_S. The superior performance of IC-RP_{AB}, IC-RRP_1, and WI-E_S for small interfaces is important; thus if NCMs are formed in the scan, a postscan re-evaluation scheme that uses these descriptors is likely to improve their ranking.

We also compared systems classified by the contribution of the electrostatic or the hydrophobic terms to the GEH score for the correct model (excluding A-Ag). The "electrostatic class" (5 E-I and 6 "other" systems) has large electrostatic scores and the "hydrophobic class" (7 E-I and 6 "other" systems) has large hydrophobic scores (exceeding 150 arbitrary score units). Both classes include many systems with large interfaces thus it is not surprising that the ROC curves and the AUC for these classes are similar to those of the large interfaces class (Fig. 4 and Table III). IC-RRP_1, IC-RP_{AB}, and WI-E_S are better descriptors for the "hydrophobic class" than for the large interfaces class in line with the higher IC-RP of aromatic and hydrophobic residues.

Are the false models favored by one descriptor the same as the false models favored by another descriptor?

Sorting of all the MolFit GEH models by the individual descriptors showed that the average rank of the best NCMs by IC-RP_{AB} (617) or IC-RRP_1 (657) is only

Table IV
The Average Percent of Common False Models Among the Highest-Ranking 3000 Models in Lists Sorted by Single Descriptors (in the upper triangle)

	GEH-Z	IC-RP_{AB}	IC-RRP_1	IC-RRP_3	WI-E_S	IC_{NC}	IC-RP_A
GEH-Z	–	32.6	34.8	29.1	31.1	36.7	30.0
IC-RP_{AB}	5.7	–	64.0	21.5	36.4	28.9	54.6
IC-RRP_1	7.9	37.1	–	23.9	41.3	28.2	48.6
IC-RRP_3	2.2	−5.4	−3.0	–	20.3	28.1	22.2
WI-E_S	4.2	9.5	14.4	−6.6	–	27.6	33.7
IC_{NC}	9.8	2.0	1.3	1.2	0.7	–	27.4
IC-RP_A	3.1	27.7	21.7	−4.7	6.8	0.5	–

For IC-RP_A the average includes only E-I and A-Ag systems. The bottom triangle shows the difference between the upper triangle and the expected percent of common false models (26.9%).

slightly better than the average rank by GEH-Z (673), and the average ranks by IC-RRP_3, WI-E_S, and IC_{NC} are clearly worse (1066, 1475, and 1170, respectively; more details are given in Supporting Information S5). A combination of descriptors is likely to distinguish more successfully between NCMs and false models than the individual descriptors, provided that all the descriptors favor NCMs but at the same time they favor different sets of false positive models (shuffling effect). We tested the shuffling effect for pairs of descriptors by intersecting the top 3000 models in lists sorted by the individual descriptors and determining the percent of common false models after each intersection (Table IV). The expected fraction of common false models if each descriptor favors a random selection of false models is 26.9 ± 1.8%. It was estimated as the ratio between the average number of false models among the top 3000 models (2993) and the average number of all the false models for the 63 systems (11,116 ± 1659). In many cases the fraction of common false models is close to or only slightly higher than the expected fraction for random selections, indicating that there is very significant shuffling of the high-ranking false models between pairs of descriptors. Significantly higher fractions of common false models are obtained for correlated descriptors, such as IC-RP_{AB} and IC-RRP_1 (64.0%). Interestingly, IC-RRP_3 clearly disfavors false models favored by most other descriptors making it potentially useful in combinations of descriptors despite its relatively low discrimination ability. Similarly, high-ranking false models selected by IC_{NC} appear to be randomly related to false models favored by the propensities and WI-E_S, which makes IC_{NC} too attractive for further testing. GEH-Z shares only up to 10% more common false models than expected with the other descriptors, hence the postscan descriptors used here are approximately orthogonal to the inscan descriptor.

Two-descriptor intersection filters

The significant shuffling of the high ranking false models together with the observation that all the descriptors favor NCMs led to the notion that intersections of lists of models sorted by the different descriptors are likely to eliminate a large fraction of false predictions. This notion was tested by repeating the 2-descriptor intersections, this time for models with descriptor values above selected thresholds (instead of the fixed number of top ranking models). The 2-descriptor intersection process is demonstrated in Figure 5, where pairs of descriptors are plotted for all the models of the E-I system 2sni. The overlaid transparent rectangles depict all the models that will be eliminated because their IC-RP_{AB}, IC-RRP_1 or WI-E_S are below the selected thresholds. The area that is not overlaid by the rectangles represents the models that are retained and it is important to note the relatively high percent of retained NCMs.

The results of two descriptor intersections of $Z1$ values are described in detail in Supporting Information S6. The average enrichment ratios achieved with single descriptor filters always exceed 1, and the enrichment ratios for the 2-descriptor intersections are always larger than the corresponding single descriptor ratios (Table S6). This result supports our notion that intersections are useful for eliminating false models. Notably, some 2-descriptor intersections are particularly effective for spe-

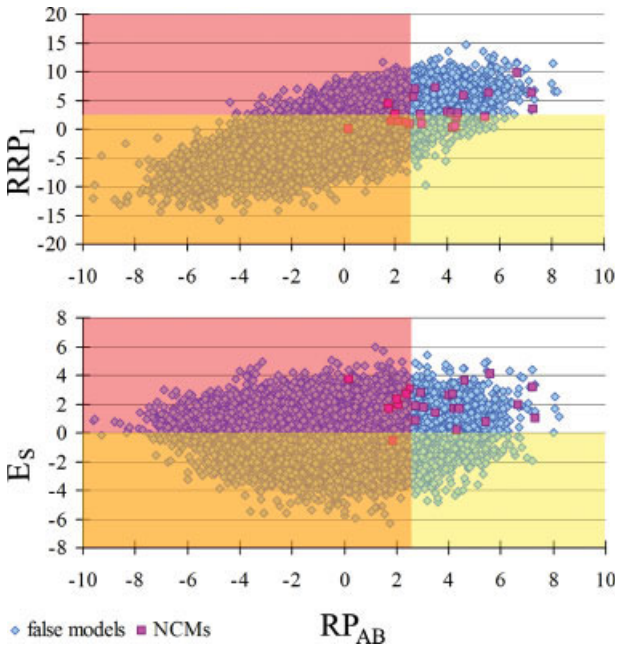

Figure 5
Demonstration of 2-descriptor intersections for the E-I system 2sni. The overlaid transparent rectangles depict models that will be eliminated because their IC-RP_{AB}, IC-RRP_1 or WI-E_S are below the selected thresholds. The area that is not overlaid by the rectangles includes the models that are retained after the intersection process and it is important to note the relatively high percent of retained NCMs in both intersections, that of the correlated descriptors IC-RRP_1 and IC-RP_{AB} and that of the noncorrelated descriptors WI-E_s and IC-RP_{AB}.

cific classes of systems, for example the enrichment ratio for the intersection of RP_{AB} and GEH-Z for E-I is 7.05 (Table S6). More importantly, some of the more effective 2-descriptor intersections for all the classes involve weakly discriminating descriptors such as IC-RRP_3 or IC_{NC}, in line with the false models shuffling results shown in Table IV.

Multi-descriptor intersections and the "soft intersection" filters

Clearly there is advantage in intersecting lists of models whose descriptor values are above preset thresholds as it eliminates a large fraction of false models and enriches the list of remaining models with NCMs. Intersections of lists sorted by more than two descriptors can be even more effective. However, our results also show that up to nine systems lose all their NCMs in the 2-descriptor intersections (Table S6) and higher dimension intersections may increase this number. We therefore tested a voting classification procedure, termed here "soft intersection" filter, in which all *M* descriptors are considered and models are retained if their descriptor values exceed the threshold for any *N* descriptors. The optimization of the class specific "soft intersection" filters and the 5-fold cross validation tests are described in Supporting Information S7. The cross validation tests indicate that the optimized class specific filters are very stable (Table S7).

The optimized intersection thresholds, listed in Table S7, are considerably higher than the preliminary thresholds used in the 2-descriptor intersections. Nevertheless, only one system out of 63 loses all its NCMs (1qa9 from the "other" class). This highlights the advantage of the "soft intersection" filter which retains models even if not all its descriptor values exceed the thresholds. We also note that the optimal thresholds are higher for the more discriminative descriptors, IC-RP_{AB}, IC-RP_A, IC-RRP_1, and GEH-*Z*, leading to the elimination of many false models without losing too many NCMs. Thus, the "soft intersection" filters retain only 758, 157, and 1218 models on average for the E-I, A-Ag, and "other" classes, respectively. Reasonable fractions of NCMs are retained for the three classes, 39.9, 11.1, and 42.9%, and significantly lower fractions of false models, 6.8, 1.2, and 10.9%, leading to average enrichment ratios of 7.6, 8.4, and 5.2. The filters discard models of all levels of accuracy. Hence, the NCMs in the unfiltered lists of models consist of 87.8, 11.8, and 0.4% of *, **, and *** accuracy models, respectively; the NCMs retained after filtering are marginally more accurate, including 87.2, 12.2, and 0.6% models with *, **, and *** accuracy, respectively.

The powerful "soft intersection" filters eliminate a large fraction of false positive models. Hence, reranking of the remaining models by their GEH-Z values improves or does not change the rank of the best NCM for 51 systems out of 62 (20/23 E-I, 8/9 A-Ag, and 23/31 "other"), and the average rank of the best NCMs improves from 683 before filtering to 208 (Tables S5 and V). Ranking of the filtered models by IC-RP_{AB} and IC-RRP_1 produces slightly better average ranks, similar to the aforementioned result for all the models.

New scoring functions based on the descriptor values

Another way to use the new descriptors is to utilize their values in class specific scoring functions, which are linear combinations of the standardized descriptor values, *Z*2. By using *Z*2 the different ranges of magnitude of the descriptor values do not affect the resultant linear combination coefficients. The optimization of the coefficients and the 5-fold cross validation of the new scoring functions are described in Supporting Information S8. The cross validation tests show that the coefficients of the new scoring functions are stable (Table S8) and that the small differences between tests do not affect the AUC for the new scoring function. We find that five descriptors contribute to the scoring functions of the E-I class (WI-E_S excluded) and the "other" class (IC_{NC} excluded) and that IC-RP_A is very dominant for the A-Ag class (Table S8 and Table V). Interestingly, IC-RRP_3 attains high coefficients in the new scoring functions for the E-I and "other" classes despite being a rather weak classifier.

The new scoring functions were optimized against lists of filtered models and their application to all the GEH models provides an independent test of their discrimination ability. We find that reranking of all the models by the new scores dramatically lowers the average rank of the best NCMs from 673 to 198 (Table S5). The ranks obtained with the new scores are higher than the GEH ranks for 43/63 systems from all three classes; at least one NCM is ranked in the range 1–50 for 34/63 systems and in the range 1–100 for 41/63 cases (Table S5).

The ROC curves for the new scoring functions (Fig. 4) and the corresponding AUC values (Table III and AUC_1

Table V
Ranking of the Best NCMs in the Filtered Lists of Models Sorted by Different Descriptors

		Distribution of the ranks of NCMs[a]		
Ranking by	Average rank	1–50	1–100	1–250
GEH-Z	208 (122,51,322)	28 (12,6,10)	40 (17,8,15)	46 (18,9,19)
IC-RP_{AB}	192 (82,71,311)	24 (12,4,8)	36 (17,6,13)	49 (21,9,19)
IC-RP_A	(84,41,-)	19 (14,5,-)	25 (16,9,-)	31 (22,9,-)
IC-RRP_1	191 (77,85,310)	32 (16,3,13)	39 (19,6,14)	47 (21,8,18)
IC-RRP_3	204 (153,52,289)	22 (8,7,7)	32 (13,7,12)	45 (21,9,15)
WI-E_S	272 (142,70,432)	18 (10,3,5)	26 (11,7,8)	41 (19,9,13)
IC_{NC}	255 (123,42,419)	22 (9,6,7)	34 (15,9,10)	44 (20,9,15)
New scoring functions	122 (45,40,205)	35 (17,5,13)	43 (20,9,14)	52 (22,9,21)

The values for classes E-I, A-Ag and "other" are given in parenthesis.
[a]Each column lists the number of systems with at least one NCM in the given ranks range and the division by classes.

in Table S7) show that the class specific new scores for E-I and A-Ag are excellent descriptors, with AUC of 0.90 and 0.85, respectively. The nonspecific new scoring function for the "other" systems has higher AUC (0.79) than any single descriptor and its Sn values are significantly higher for 1-Sp >0.07 (Fig. 4).

Combining the "soft intersection" filters with the new scoring functions

The "soft intersection" filters and the re-evaluation by the new scoring functions can be combined, improving (or not changing) the ranking of the best NCM for 46/62 systems and worsening the rank for only 16 cases. The positive changes in rank are much larger than the negative changes leading to a very significant improvement of the average rank of the best NCMs to 122 (Table V). The rank changes for all the systems are demonstrated in Figure 6 and listed in Table S1. In some cases the rank improvement is dramatic; for example, the rank of the best NCM for system 1ezu improves from 2436–2490 to 152–153 and for system 1gp2 it improves from 1150–1178 to 4 (Table S1). It appears that even after the extensive elimination of models by the "soft intersection" filters, which improve the average rank from 683 to 208 (for 62 systems), the re-evaluation by the new scoring functions further improves the average rank to 122. For 84% of the systems (52/62) an NCM is ranked <250 and 70% of the systems have at least one NCM ranked <100 (Table V); for 18 systems at least one NCM is ranked in the range 1–10.

To further demonstrate the effectiveness of the combined filtering and re-evaluation procedures, we recalculated AUC values for all the GEH models giving a score of 0 to models discarded by the "soft intersection" filter (AUC_2 in Supporting Information S8). The resultant AUC are very high, 0.96, 0.99, and 0.93 for the E-I, A-Ag, and "other" classes.

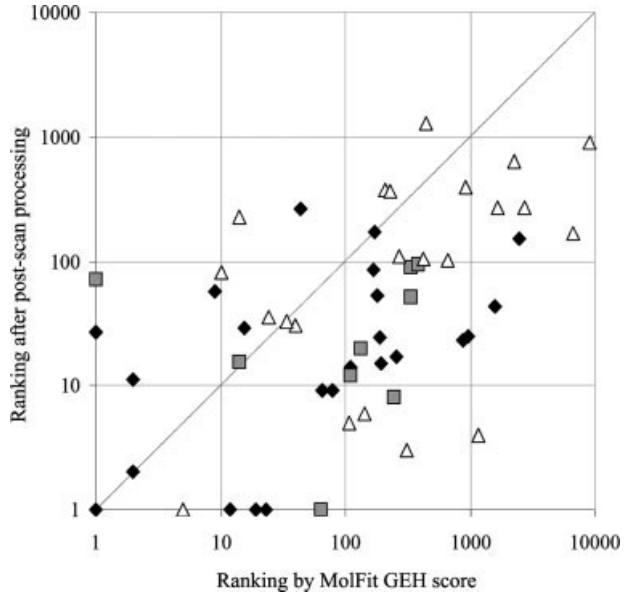

Figure 6

Comparison of the ranks of the best NCMs before and after postscan processing for 62 systems. The points below the diagonal represent all the systems for which the post-scan processing improved the rank of the best NCM (black diamonds for E-I, gray squares for A-Ag systems and empty triangles for "other" systems). Note the log/log scale of the best NCMs ranking and the narrower range of ranks after the post scan processing.

Application to CAPRI targets

The new postscan processing procedures were tested on 23 CAPRI targets[69–71] that are not included in our database and for which NCMs were formed in the scan (excluding homomultimers; Table VI). The procedures were applied to lists of models obtained with MolFit during the prediction periods. Some of these lists are based on only geometric or weighted geometric scans and some are the result of weighted scans followed by GEH intersection and additional filtering according to external experimental and bioinformatics' data. Hence, these are not optimal lists of models for our postscan processes, which are optimized for lists of GEH models. Application of the new procedures to such lists provides both, validation of their effectiveness and a test of their robustness.

Five of the CAPRI targets in Table VI are E-I systems: T15,[80] T16,[81] T17,[81] T18,[82] and T32.[90] Application of the "soft intersection" filter to the geometric scan results for the first four targets is very effective, reducing the number of putative models from 22,200 to less than 1000 and improving very significantly the ranks of the NCMs. In three of the four cases ranking by the new scoring function further improves the rank of the best NCM. Particularly interesting is T18, where the best NCM is ranked low by the geometric score[91] probably due to a movement of loop 113-128 that narrows the active site in the unbound enzyme causing clashes with the inhibitor. The new scoring function ranks the best NCM 9. For T15 the side chains were "shaved" and then remodeled.[70] Our remodeling was rough and we made no attempt to optimize it.[91] The rough shape of the molecules was nevertheless tolerated in the MolFit scan that produced 17 NCMs and is also tolerated in the new post scan process, which improves the rank of the NCM. Notably, experimental information is not considered in any of these four tests; inclusion of such information will most likely improve the ranks of NCMs. Our E-I filter discards all 24 NCMs for the protease-inhibitor target T32 (interface A:C of 3bx1). Detailed check indicates that the GEH-Z, IC-RRP_1, and WI-E_s values for the NCMs of T32 are high whereas the IC-RP_{AB}, IC-RRP_3, and IC_{NC} values for most of the NCMs are low; similar descriptor values are obtained also for the experimental interface.

Table VI
Application of the Postscan Processing Procedures to CAPRI Targets

		MolFit scan and filtering results		After ``soft intersection'' filtering and re-evaluation by the new scoring functions				
Target	Type of complex and scan[a]	No. of models (NCMs)	Rank[b]	Type of filter	No. of models (NCMs)	Rank by the MolFit score[c]	Rank by the new score	Enrichment ratio
T02:VP6/antibody[d]	B-U/wG	25320 (518)	28	A-Ag	613 (87)	4	82	7.9
T03:HA/antibody[72]	B-U/wG	12660 (268)	1	A-Ag	258 (4)	2	33	0.7
T04:α-amylase/cameloid antibody[73]	B-U/G	17520 (19)	732	E-I	782 (4)	278	530	4.7
				A-Ag	21 (0)	–	–	0.0
				``Other''	1487 (4)	401	761	2.5
T05:α-amylase/cameloid antibody[73]	B-U/G	17520 (6)	2144	E-I	647 (1)	440	96	4.5
				A-Ag	16 (0)	–	–	0.0
				``Other''	1327 (2)	389	138	4.4
T06:α-amylase/cameloid antibody[73]	B-U/G	8760 (17)	6	E-I	348 (13)	1	2	20.0
				A-Ag	14 (2)	8	7	85.7
				``Other''	727 (16)	3	4	11.6
T07:TCRβ/toxin[74]	U-U/G	8760 (30)	962	``Other''	1517 (6)	4	13	1.2
T08: nidogen/laminin[75]	B-U/G	17520 (15)	183	``Other''	1439 (8)	31	47	6.5
	B-U/wG	17520 (45)	3		1428 (24)	3	15	6.3
T12: cohesin/dockerin[76,77]	B-U/G	17520 (213)[e]	16	``Other''	2065 (54)	3	157	2.2
T13: SAG1/antibody[78]	B-U/G	17520 (33)	1590	A-Ag	168 (1)	102	131	3.2
	B-U/wG	17520 (102)	2		88 (1)	23	80	2.0
T14:phosphatase/MYPT1[79]	M-B/G	22200 (9)	43	``Other''	2825 (9)	19	8	7.9
T15:colicinD/ImmD[80]	Bm-Bm/G	22200 (17)	3045	E-I	846 (1)	808	796	1.5
T16:xylanase(GH10)/XIP-I[81]	M-U/G	22200 (9)	2881	E-I	812 (4)	396	22	12.2
T17:xylanase(GH11)/XIP-I[81]	M-U/G	22200 (12)	900	E-I	915 (8)	118	190	16.3
T18:xylanase(GH11)/TAXI-I[82]	B-U/G	22200 (14)	505	E-I	672 (8)	79	9	19.1
T19:prion/antibody[83]	M-B/G	22200 (29)	565	A-Ag	26 (1)	23	1	30.6
T21:Orc1/Sir1[84]	U-U/fGEH	1256 (10)	64	``Other''	110 (2)	94	37	2.3
T22:U5-15k/U5-52k[85]	U-U/GEH	9144 (10)	89	``Other''	1211 (0)	–	–	0.0
T25:Arf1/ARHGap21[86]	U-B/fGEH	454 (10)	1	``Other''	55 (7)	2	1	5.5
T26:TolB/Pal[87]	U-U/fGEH	550 (41)	1	``Other''	57 (7)	1	3	1.7
T27:E225K/Ubc9[f]	U-U/fGEH	110 (6)	3	``Other''	7 (4)	1	2	23.0
T29:Trm8/Trm82[88]	U-B/fGEH	138 (1)	24	``Other''	26 (1)	7	4	5.5
T30:PLXNB1/RND1[89]	U-U/fGEH	471 (1)	87	``Other''	60 (1)	26	27	8.0
T32:savinase/BASI[90]	U-U/fGEH	1011 (24)	140	E-I	44 (0)	–	–	0.0
Average for G and wG targets		18420 (82.6)	894.7		1020.5 (15.1)	175.3	157.2	9.1
Average for GEH and fGEH targets[g]		496.5 (11.5)	30.0		52.5 (3.7)	21.8	12.3	7.7

[a]This column lists the type of target and the type of MolFit scan/filter used in the prediction process.
[b]Rank of the best NCM.
[c]Different scoring functions were used here, according to the prediction procedure, however in every case these are the same scoring functions as in column 4.
[d]The experimental structure of this CAPRI target, contributed by Vaney MC, and Rey F, is not published.
[e]The two possible binding modes of the dockerin are included.
[f]The experimental structure of this CAPRI target was deposited in the Protein Data Bank, code 2o25, by Walker, J.R., Avvakumov, G.V., Xue, S., Newman, E.M., Mackenzie, F., Weigelt, J., Sundstrom, M., Arrowsmith, C.H., Edwards, A.M., Bochkarev, A. and Dhe-Paganon, S.
[g]Six cases for which at least one NCM was retained after application of the post scan filter were included in these averages.
U – unbound, B – bound, M – modeled, Bm – bound with modeled side chains, G – geometric scan, wG – geometric scan with weighting based on external data,[11] GEH – geometric-electrostatic-hydrophobic intersection,[20] fGEH – intersection of weighted geometric, electrostatic and hydrophobic scan results followed by filtering of the models based on external data.

The new A-Ag "soft intersection" filter is very effective for the geometric scan results for two A-Ag targets, T13[78] and T19.[83] In both cases the filter retains less than 200 models and the ranks of the NCMs are improved (Table VI). The "soft intersection" filter successfully discards a large number of models also from the weighted geometric lists for targets T02, T03,[72] and T13,[78] but the new A-Ag scoring function, which is dominated by the descriptor IC-RP_A, ranks NCMs lower than the weighted geometric score. This result is not surprising. The weighted-geometric scans involve upweighting of interactions made by the antibody complementarity determining regions (CDRs) leading to formation of more models (NCMs and false) in which these regions contact the antigen. The CDRs are enriched with residues with high propensity to be in the IC leading to high IC-RP_A values for the NCMs but also for false models thus weakening the distinction between them.

Eleven of the CAPRI targets belong to the "other" class. In four of these cases (targets T07,[74] T08,[75] T12,[76,77] and T14[79]) the new post scan procedure was applied to models produced by only geometric scans. The "other" filter efficiently eliminated most of the false models enriching the remaining models with NCMs;

however the new scoring function ranked the best NCM better than the geometric score only for T14. In seven other cases the post scan procedure was applied to GEH models or to GEH models filtered by external data (targets T21,[84] T22,[85] T25,[86] T26,[87] T27, T29,[88] and T30[89]); both the "soft intersection" filter and the new scoring function proved to be very effective in these cases. The filtered GEH lists of models are short, including between 1256 and 110 models. Nevertheless the "soft intersection" filter reduces the number of models in each list considerably, leaving 110 models in the longest list. The ranks of the NCMs were initially high for T25, T26, and T27, between 1 and 3, and we note that the postscan processing did not change the ranking in these cases by more than ± 2. In three other cases (T21, T29, and T30) the combination of "soft intersection" filtering and re-evaluation by the new scoring function improved the rank significantly, from 64, 24, and 87 to 37, 4, and 27 for T21, T29, and T30, respectively. Notably, the interfaces of T07 and T21 are small and more difficult to predict and the success of the new scoring function is therefore important. In one case, T22, the "soft intersection" filter eliminates all the NCMs, which have consistently low IC-RP_{AB} and IC-RRP_1 and negative WI-E_s, as does the experimental interface.

Targets T04,[73] T05,[73] and T06[73] are interesting as they involve interaction of an enzyme (α-amylase) and a cameloid antibody and it cannot be decided in advance if the antibody acts as an inhibitor (hence an E-I target), as an antibody (hence an A-Ag target) or if the interaction does not relate to either the catalytic site of the enzyme or the CDRs of the antibody (hence an "other" target). We therefore applied each of the three different post scan procedures to the geometric scan results for these targets. It appears that the A-Ag "soft intersection" filter is not adequate for cameloid antibodies as it eliminates most or all the NCMs. This result is sensible in view of the smaller number of antibody CDR residues involved in these interactions (a cameloid antibody consists of only one antigen binding domain cf. with two such domains in other antibodies) and the involvement of other parts of the antibody in antigen binding. In contrast, postscan processing with the "other" function retains NCMs in every case and ranks the best NCM similarly or better than the geometric score. The E-I postscan procedure produces marginally better results than the "other" postscan procedure for target T06, where the cameloid antibody binds in the active site of the enzyme.

The analysis of CAPRI targets shows that application of the "soft intersection" filter to lists of models already filtered according to external data relevant to the specific interaction can be very beneficial, as it significantly reduces the number of false models. For example, it retained only seven models for target T27 (out of 110), four of which are NCMs. We note that the "soft intersection" filtering is effective also for models produced in geometric or weighted-geometric scans reducing dramatically the number of models yet retaining some NCMs. The new scoring functions appear to be more effective for filtered GEH models than for weighted geometric models. They improve or do not change significantly the ranking of the best NCM for 13 CAPRI targets.

DISCUSSION

Computational protein-protein docking is important for gaining better understanding of protein interactions, their role in diseases, and for drug development. Although prediction of protein-protein complexes is a difficult task due to the enormous number of possible binding configurations, NCMs are very often produced in the docking scan. However, the scoring functions used in the scan do not always rank high the NCMs, in particular in the absence of external information on the interaction of interest. We present in this study, postscan filtering and re-evaluation procedures that employ interface descriptors that are not used in the docking scan. These descriptors are based only on the structures of the molecules and do not use external data. Unlike in many other studies where the attributes of the interface are optimized to distinguish between native interfaces and false models, we optimized the parameters of our procedures by comparing NCMs and false models because we surmised that NCMs may be too different from native structures. Indeed, the new procedures very significantly enhance the discrimination between NCMs and false models.

Interface core properties

We show here that RP and RRP calculated for the IC are discriminative descriptors. The performance of the propensity descriptors is only moderate when the strict definition of the IC (totally buried interface atoms[42]) is employed and it greatly improves when atoms with exposed surface up to 5 $Å^2$ are included. This result is not entirely unexpected. Thus, the interface of a NCM in unbound rigid-body docking is not as compact as in the native complex due to conformation differences between the bound and unbound structures; the slightly relaxed IC definition compensates for this mismatch. The ICp program allows matching of the definition of the IC atoms (and hence residues) to each descriptor. Accordingly, the performance of IC-RP_{AB} is best when a more stringent definition is employed ($EX_{max} \leq 5$ $Å^2$) in line with the fact that the RP values were derived for the IC.[42] The RRP tables used here were derived for whole interfaces[27,29,30] and for these descriptors more relaxed IC definitions ($EX_{max} \leq 10$ $Å^2$ or 15 $Å^2$) perform slightly better. Notably, this is a small range of EX_{max}; for comparison the surface area of a solvated oxygen atom (a sphere of radius 2.8 Å) is approximately 100 $Å^2$.

In contrast to the propensity descriptors, the solvation energy tested in this study, although only moderately discriminative in general, clearly performs better when calculated for the whole interface. This result cannot be attributed to the nature of the preferred IC residues, which include many hydrophobic residues with positive solvation energies. It may reflect a statistical aberration, which is larger for the smaller number of contributing atoms in the IC.

The IC is a relatively new concept and the IC-RP was used in a scoring function only in one very recent study. Bernauer *et al.*[40] defined the interface and the IC using Voronoï tessellation and tested physical contacts. They found that the size of the interface and its compactness are good discriminators between native and false interfaces; IC-RP and IC-RRP (defined for residue categories) are useful though weaker discriminators. The IC in the study by Bernauer *et al.* is different from the IC in this study not only in the way it is calculated but also in the strictness of the definition; we allow slightly different definitions for each individual descriptor.

Class dependencies

Another important issue that this study highlights is the differences in relevance and performance of some descriptors for classes of systems divided by their biological role. The performance of the per-molecule descriptors IC-RP_A and IC-RP_B is particularly interesting. For E-I systems the discrimination ability of IC-RP_A (for the enzyme) and IC-RP_B (for the inhibitor) is similar. This result is in-line with the notion that both the enzyme and the inhibitor underwent evolution to have matching interaction sites. In contrast, the per-molecule IC-RP calculations for the A-Ag systems show that while the ICs of antibodies include high proportion of residues with high IC-RP, the antigens do not. Furthermore, the ROC curve for IC-RP_B (IC-RP for antigens) is below the diagonal, indicating that antigen interfaces are enriched with residues that have negative IC propensities. Indeed Asn and Lys are abundant in the antigen interfaces tested here, in agreement with the observation that the most frequent residues in antigenic epitopes are Asn, Pro, Arg, and Lys[67] and with the calculations of Jackson[47] who showed that Asn, Arg, Lys, and Asp contribute most to the antigen interaction energies. Asn, Pro, and Lys are under-represented in the IC with Lys being the rarest IC residue.[42] Interestingly, we find that the antigenic site propensities[67] do not discriminate between NCMs and false models of A-Ag systems probably because they have similar trends to surface propensities.[68] These results strongly suggest that antigen interfaces should be excluded from the derivation of RP and RRP tables as they do not carry distinct interface information.

Differences between E-I and A-Ag systems were noted in several previous studies.[12–14,47] Li *et al.*[52] used class specific linear combinations of descriptors and found that for E-I systems electrostatics and RP are most useful whereas for A-Ag systems electrostatics, RP and atomic contact energies have similar effect. Class specific scoring functions were presented also by Martin and Schomburg.[51] They too found that an optimized scoring function that includes many descriptors is efficient for re-evaluating docking models. Interestingly, they found that RP derived specifically for A-Ag systems dominates the combined score for this class.

The class analysis in this study brought out additional insights. Thus, IC-RP_{AB}, IC-RRP_1, and WI-E_S are significantly more discriminative for small interfaces than for large interfaces. This is important in view of the dominant roles of geometric terms in different scoring functions that hampers the identification of small interfaces.[32,61]

RP *versus* RRP descriptors

A somewhat bewildering result of our computations is the large difference between the discrimination ability of RP and RRP. RP is highly discriminative both between binding and nonbinding interface regions and between proteins with different biological roles. RP values represent the energy change upon burial of a residue in the interface. It is not surprising therefore that the more abundant residues in ICs of nonobligatory heterodimers (Trp, Tyr, Met, Phe, Cys, His, Ile, Leu, and Arg)[42] are mostly hydrophobic, including the large aromatic residues, whose desolvation is energetically beneficial. Presumably, evolutionary forces selected these residues for maintaining protein-protein interactions also because they are generally large and able to make numerous contacts with the binding partner; most of them can participate in different types of interactions - structurally nonspecific van der Waals contacts and orientation specific hydrogen bonds.

In this study, RRPs are weak discriminators, hence, only RRP_1, which also carries RP information, distinguishes well between NCMs and false models, while RRP_2 and RRP_3 are considerably less discriminative. RRP_2 was derived using a database of homo- and heterodimers (~2:1 ratio) and a tolerant measure for detecting contacting residue pairs (Cβ-Cβ distances). RRP_3 is based on physical residue-residue contacts in hetero-complexes, making it perhaps more adequate for our IC-RRP calculations. These differences may explain the weak correlation between the IC-RRP_2 and IC-RRP_3 values ($r = 0.36$, see Supporting Information S4), which reflects the very weak correlation between the propensity tables (data not shown). For example, contacts between aromatic and hydrophobic-aliphatic residues are favored more in the RRP_3 table than in the RRP_2 table, which may explain the slightly better discrimination ability of IC-RRP_3.

The weak discrimination ability of the RRP descriptors is surprising as the RRP tables reflect interaction prefer-

ences. Possibly statistical fluctuations are larger when 2-body interactions (between pairs of residues) are counted than when single residues are considered. Defining residue contacts via atom-atom distances may also introduce a considerable error in the derivation of the residue-residue frequency tables and in the calculation of the RRP values for interfaces. Also, the low discrimination ability of RRP may be due to the fact that some of the features that it represents are considered in the GEH score. Finally, RRP derived for only the IC may be more discriminative, as is the IC-RP compared with whole interface RP. An attempt in this direction,[40] in which pair frequencies for residue categories were derived using only physically contacting pairs, showed strong preference for interactions involving hydrophobic and aromatic residues. Our calculations show that despite their relatively weak discrimination ability, RRP carry different information than the RP leading to dissimilar selection of high ranking false models by IC-RP_{AB} and IC-RRP_3 (Table IV) and to elimination of many false models in the 2-descritptor intersections (Supporting Information S6) and the multi-descriptor intersections.

Combining descriptors

Although IC-RRP_1 and especially IC-RP_{AB} are good discriminators, neither of them is significantly better than the GEH score for ranking NCMs. We therefore combined the descriptors. In general, such combinations are likely to be more effective if the individual descriptors are orthogonal. Most of the descriptors tested in this study are indeed noncorrelated or very weakly correlated. It appears however that even correlated descriptors should be considered in the combination and IC-RP_{AB} and IC-RRP_1 provide a good example. They are significantly correlated ($r = 0.72$) and their ROC curves are very similar for classes of systems with defined biological role (the E-I and A-Ag classes; Fig. 4), almost superposed. However, when the class is not uniform in terms of the biological role ("other" and classes selected according to the native interface size or chemical nature) the ROC curves of these two descriptors diverge, revealing that each of them carries additional independent information and that the variation between the systems affects them differently.

We combined the descriptors in class specific postscan processing procedures by (1) using the descriptors as Boolean yes/no classifiers and (2) using the standardized descriptors in new scoring functions. The "soft intersection" filter presented here was not used before for filtering docking models; it requires that the descriptors of a given model exceed N out of M thresholds. Naturally there is a tradeoff between the thresholds used in the filter and the number of retained models. The combination of relatively high thresholds for some descriptors (hence retaining fewer models), with the "soft intersection" process produces significantly shorter lists of models and much higher ranks of the best NCMs (average of 208 *vs.* 683 for 62 systems). Evidently many false models are eliminated in the process, including many false positives.

The new, class specific, scoring functions were applied to all the MolFit GEH models and to the filtered models. In this process the number of models does not change but the ranks of the models change according to the new scores. When applied to all the GEH models, the new scoring functions improve the average rank of the best NCMs from 673 to 198 (for 63 systems). Interestingly, applying the new scoring function to the filtered lists of models further improves the average rank of the best NCMs from 208 to 122. This additional rank improvement is a result of the different use of the descriptor values in the "soft intersection" filtering and the re-evaluation procedures, as the first procedure eliminates false models (including false positive), whereas the second procedure highlights NCMs. In this study we focus on the performance of the new post scan processes, which rely only on the structures of the molecules and properties derived from the structures. The ranks of NCMs can be further elevated by employing algorithms that cluster structurally similar models (false and NCMs) and eliminate replicates, or by using external information as demonstrated for several CAPRI targets.

The performance of our postscan procedures can be compared with other such procedures, which make no use of external data and do not modify the models via energy minimization. For example, the scoring function developed by Pierce and Weng,[38] which is a weighted sum of van der Waals attractive and repulsive terms, short- and long-range electrostatic energy terms and desolvation, and the pyDock re-evaluation procedure, which employs electrostatic and desolvation terms.[39] Fifty-two systems are common to our dataset and the dataset tested by Pierce and Weng.[38] Our procedure produces an average rank of 106 for these systems compared with an average rank of 2718 produced by ZDOCK2.3+ZRANK; in 21 cases we rank an NCM among the top 20 models, compared with 18 such cases in their study. Also, Pierce and Weng did not find NCMs (hits) among the top 2000 models for 9 of the 52 systems, whereas in this study only one system did not have NCMs among the 36–1802 models retained by the "soft intersection" filter for the same set of systems. The average rank of the best NCMs for the 49 systems common to our study and the pyDock study, is 104 in this study compared with an average rank of 600 by pyDock.[39]

Murphy *et al.*[92] applied several filters, such as atomic contact potentials, RRP and electrostatics, to models generated by rigid body docking. The 10 systems tested in their study are almost all from the E-I class. The average percent of retained NCMs (estimated based on their Fig. 2 and Table IV) is ~47%, similar to the corresponding value in this study, 40%, for 23 E-I systems.

As mentioned above, subgroup specific scoring functions were employed by Li *et al.*[52] and by Martin and Schomburg.[51] The average ranks of the best NCMs obtained in this study, 45, 40, and 205 for the E-I, A-Ag, and "other" classes, respectively, are significantly better than the values published by Martin and Schomburg,[51] 84, 341, and 568, for the same classes of systems. The number of systems with at least one NCM among the 10 top ranking models is 7/23, 2/9, and 9/31 for E-I, A-Ag, and "other" (18/63 for all tested systems) in our study, in which only unbound-unbound docking cases are considered, compared with 11/22, 4/21, and 1/29 (16/72) in the study of Martin and Schomburg[51] and 1/6, 6/19, and 6/15 (13/40) in the study by Li *et al.*,[52] where bound-unbound A-Ag cases were included.

The post scan processing procedures presented here prove to be very powerful when external information regarding the interaction of interest is unavailable and the docking results rely only on the physicochemical properties of the component molecules (Fig. 6). They are useful also when such information exists and was used in the production of the docking models, as demonstrated for several CAPRI targets. Our results for the CAPRI targets can be compared only generally with the results obtained by Bernauer *et al.*[40] because different definitions of NCMs were used in the two studies. The average rank of the rank1 solutions for T12-T14 and T16-T19 in their study, obtained by rescoring models produced by DOCK, is 25.6 (Table 1 in[40]; T15 was excluded from this comparison because they quote *Fnat* = 0 for this target), better than the average rank for the same targets in our study: 105.7 after application of the "soft intersection" filters and 74 after re-evaluation by our new scoring functions. However, for most of these targets the NCMs listed by Bernaur *et al.*[40] may be unacceptable by the CAPRI assessment criteria employed here.

In summary, we present in this study new concepts such as the per-molecule interface-core propensities (IC-RP_A and IC-RP_B) and the shuffling of false models, and a new tool - the "soft intersection" filter. The new concepts can be used in the future for development of additional postscan processing tools that use more descriptors and pertain to other classes of systems. Our class specific postscan processing combines very efficient elimination of false models by a voting classification procedure, the "soft intersections", and accentuation of NCMs by new scoring functions. Although these procedures are optimized and tested here for models produced with our docking algorithm MolFit, they are likely to be effective also for models produced with other rigid body docking programs.

ACKNOWLEDGMENTS

The authors thank Professor Ephraim Katchalski-Katzir for his support throughout this study. They also thank Amnon Horovitz, Lia Addadi, Shai Rosenwald and members of our group, Avraham Ben-Shimon, Ziv Frankenstein and Maya Kahan for helpful discussions. MolFit can be freely downloaded from our website http://www.weizmann.ac.il/Chemical_Research_Support/molfit/.

REFERENCES

1. Janin J, Seraphin B. Genome-wide studies of protein-protein interaction. Curr Opin Struct Biol 2003;13:383–388.
2. Bahadur RP, Zacharias M. The interface of protein-protein complexes: analysis of contacts and prediction of interactions. Cell Mol Life Sci 2008;65:1059–1072.
3. Bonvin AM. Flexible protein-protein docking. Curr Opin Struct Biol 2006;16:194–200.
4. Gray JJ. High-resolution protein-protein docking. Curr Opin Struct Biol 2006;16:183–193.
5. Eisenstein M, Katchalski-Katzir E. On proteins, grids, correlations, and docking. CR Biol 2004;327:409–420.
6. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
7. Wodak SJ, Janin J. Structural basis of macromolecular recognition. Adv Protein Chem 2002;61:9–73.
8. Sternberg MJ, Gabb HA, Jackson RM. Predictive docking of protein-protein and protein-DNA complexes. Curr Opin Struct Biol 1998;8:250–256.
9. Janin J. Protein-protein recognition. Prog Biophys Mol Biol 1995; 64:145–166.
10. van Dijk AD, Boelens R, Bonvin AM. Data-driven docking for the study of biomolecular complexes. FEBS J 2005;272:293–312.
11. Ben-Zeev E, Eisenstein M. Weighted geometric docking: incorporating external information in the rotation-translation scan. Proteins 2003;52:24–27.
12. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
13. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
14. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
15. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004; 336:943–955.
16. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990;265:16027–16030.
17. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
18. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. J Mol Biol 1993;234:946–950.
19. Cazals F, Proust F, Bahadur RP, Janin J. Revisiting the Voronoi description of protein-protein interfaces. Protein Sci 2006;15:2082–2092.
20. Berchanski A, Shapira B, Eisenstein M. Hydrophobic complementarity in protein-protein docking. Proteins 2004;351:309–326.
21. De S, Krishnadev O, Srinivasan N, Rekha N. Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. BMC Struct Biol 2005;5:15.
22. McCoy AJ, Chandana Epa V, Colman PM. Electrostatic complementarity at protein/protein interfaces. J Mol Biol 1997;268:570–584.
23. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol 2000;10:153–159.
24. Heifetz A, Katchalski-Katzir E, Eisenstein M. Electrostatics in protein-protein docking. Protein Sci 2002;11:571–587.

25. Bordner AJ, Abagyan R. Statistical analysis and prediction of protein-protein interfaces. Proteins 2005;60:353–366.
26. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng 2000;13:77–82.
27. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001; 43:89–102.
28. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003;325:377–387.
29. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J 2003;84:1895–1901.
30. Anashkina A, Kuznetsov E, Esipova N, Tumanyan V. Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces. Proteins 2007;67:1060–1077.
31. Yan CWF, Jernigan RL, Dobbs D, Honavar V. Characterization of protein-protein interfaces. Protein J 2008;27:59–70.
32. Kowalsman N, Eisenstein M. Inherent limitations in protein-protein docking procedures. Bioinformatics 2007;23:421–426.
33. Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35:364–373.
34. Palma PN, Krippahl L, Wampler JE, Moura JJG. BiGGER: A new (soft) docking algorithm for predicting protein interactions. Proteins 2000;39:372–384.
35. Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins 2002;47:281–294.
36. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331:281–299.
37. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 2006;65: 392–406.
38. Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 2007;67:1078–1086.
39. Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 2007;68:503–515.
40. Bernauer J, Aze J, Janin J, Poupon A. A new protein-protein docking scoring function based on interface residue properties. Bioinformatics 2007;23:555–562.
41. Smith GR, Sternberg MJ, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol 2005;347:1077–1101.
42. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47:334–343.
43. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003;53:708–719.
44. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci USA 2005;102:15447–15452.
45. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
46. Moreira IS, Fernandes PA, Ramos MJ. Hot spots-A review of the protein-protein interface determinant amino-acid residues. Proteins 2007;68:803–812.
47. Jackson RM. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. Protein Sci 1999;8:603–613.
48. Nooren IM, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991–1018.
49. Nooren IM, Thornton JM. Diversity of protein-protein interactions. EMBO J 2003;22:3486–3492.
50. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA 2005;102:10930–10935.
51. Martin O, Schomburg D. Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. Proteins 2008;70:1367–1378.
52. Li CH, Ma XH, Shen LZ, Chang S, Chen WZ, Wang CX. Complex-type-dependent scoring functions in protein-protein docking. Biophys Chem 2007;129:1–10.
53. Muller W, Sticht H. A protein-specifically adapted scoring function for the reranking of docking solutions. Proteins 2007;67:98–111.
54. Janin J. Welcome to CAPRI: a critical assessment of predicted interactions. Proteins 2002;47:257.
55. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein-protein docking benchmark 2.0: an update. Proteins 2005;60:214–216.
56. Mendez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 2003;52:51–67.
57. Mendez R, Leplae R, Lensink MF, Wodak SJ. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins 2005;60:150–169.
58. Hubbard SJ, Thornton JM. 'NACCESS', Computer Program. London: Department of Biochemistry and Molecular Biology, University College London; 1993.
59. Cummings MD, Hart TN, Read RJ. Atomic solvation parameters in the analysis of protein-protein docking results. Protein Sci 1995;4: 2087–2099.
60. Vuk M, Curk T. ROC curve, lift chart and calibration plot. Adv Methodol Stat (Metodoloski zvezki) 2006;3:89–108.
61. Vajda S. Classification of protein complexes based on docking difficulty. Proteins 2005;60:176–180.
62. Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 2006;7:27.
63. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–797.
64. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. Bioinformatics 2008;24:652–658.
65. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
66. Burgoyne NJ, Jackson RM. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. Bioinformatics 2006;22:1335–1342.
67. Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci 2006;15:2558–2567.
68. Negi SS, Braun W. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. J Mol Model 2007; 13:1157–1167.
69. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a critical assessment of predicted interactions. Proteins 2003;52:2–9.
70. Janin J. The targets of CAPRI rounds 3–5. Proteins 2005;60:170–175.
71. Janin J. The targets of CAPRI rounds 6–12. Proteins 2007;69:699–703.
72. Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, Knossow M. An antibody that prevents the hemagglutinin low pH fusogenic transition. Virology 2002;294:70–74.
73. Desmyter A, Spinelli S, Payan F, Lauwereys M, Wyns L, Muyldermans S, Cambillau C. Three camelid VHH domains in complex with porcine pancreatic alpha-amylase. Inhibition and versatility of binding topology. J Biol Chem 2002;277:23645–23650.

74. Sundberg EJ, Li H, Llera AS, McCormick JK, Tormo J, Schlievert PM, Karjalainen K, Mariuzza RA. Structures of two streptococcal superantigens bound to TCR beta chains reveal diversity in the architecture of T cell signaling complexes. Structure 2002;10:687–699.
75. Takagi J, Yang Y, Liu JH, Wang JH, Springer TA. Complex between nidogen and laminin fragments reveals a paradigmatic beta-propeller interface. Nature 2003;424:969–974.
76. Carvalho AL, Dias FM, Prates JA, Nagy T, Gilbert HJ, Davies GJ, Ferreira LM, Romao MJ, Fontes CM. Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. Proc Natl Acad Sci USA 2003;100:13809–13814.
77. Carvalho AL, Dias FM, Nagy T, Prates JA, Proctor MR, Smith N, Bayer EA, Davies GJ, Ferreira LM, Romao MJ, Fontes CM, Gilbert HJ. Evidence for a dual binding mode of dockerin modules to cohesins. Proc Natl Acad Sci USA 2007;104:3089–3094.
78. Graille M, Stura EA, Bossus M, Muller BH, Letourneur O, Battail-Poirot N, Sibai G, Gauthier M, Rolland D, Le Du MH, Ducancel F. Crystal structure of the complex between the monomeric form of *Toxoplasma gondii* surface antigen 1 (SAG1) and a monoclonal antibody that mimics the human immune response. J Mol Biol 2005; 354:447–458.
79. Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R. Structural basis of protein phosphatase 1 regulation. Nature 2004;429:780–784.
80. Graille M, Mora L, Buckingham RH, van Tilbeurgh H, de Zamaroczy M. Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. EMBO J 2004;23:1474–1482.
81. Payan F, Leone P, Porciero S, Furniss C, Tahir T, Williamson G, Durand A, Manzanares P, Gilbert HJ, Juge N, Roussel A. The dual nature of the wheat xylanase protein inhibitor XIP-I: structural basis for the inhibition of family 10 and family 11 xylanases. J Biol Chem 2004;279:36029–36037.
82. Sansen S, De Ranter CJ, Gebruers K, Brijs K, Courtin CM, Delcour JA, Rabijns A. Structural basis for inhibition of *Aspergillus niger* xylanase by triticum aestivum xylanase inhibitor-I. J Biol Chem 2004;279:36022–36028.
83. Eghiaian F, Grosclaude J, Lesceu S, Debey P, Doublet B, Treguer E, Rezaei H, Knossow M. Insight into the PrPC–>PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. Proc Natl Acad Sci USA 2004;101:10254–10259.
84. Hou Z, Bernstein DA, Fox CA, Keck JL. Structural basis of the Sir1-origin recognition complex interaction in transcriptional silencing. Proc Natl Acad Sci USA 2005;102:8489–8494.
85. Nielsen TK, Liu S, Luhrmann R, Ficner R. Structural basis for the bifunctionality of the U5 snRNP 52K protein (CD2BP2). J Mol Biol 2007;369:902–908.
86. Menetrey J, Perderiset M, Cicolari J, Dubois T, Elkhatib N, El Khadali F, Franco M, Chavrier P, Houdusse A. Structural basis for ARF1-mediated recruitment of ARHGAP21 to Golgi membranes. EMBO J 2007;26:1953–1962.
87. Bonsor DA, Grishkovskaya I, Dodson EJ, Kleanthous C. Molecular mimicry enables competitive recruitment by a natively disordered protein. J Am Chem Soc 2007;129:4800–4807.
88. Leulliot N, Chaillet M, Durand D, Ulryck N, Blondeau K, van Tilbeurgh H. Structure of the yeast tRNA m7G methylation complex. Structure 2008;16:52–61.
89. Tong Y, Chugha P, Hota PK, Alviani RS, Li M, Tempel W, Shen L, Park HW, Buck M. Binding of Rac1, Rnd1, and Rho D to a novel Rho GTPase interaction motif destabilizes dimerization of the plexin-B1 effector domain. J Biol Chem 2007;282:37215–37224.
90. Micheelsen PO, Vevodova J, De Maria L, Ostergaard PR, Friis EP, Wilson K, Skjot M. Structural and mutational analyses of the interaction between the barley alpha-amylase/subtilisin inhibitor and the subtilisin savinase reveal a novel mode of inhibition. J Mol Biol 2008;380:681–690.
91. Ben-Zeev E, Kowalsman N, Ben-Shimon A, Segal D, Atarot T, Noivirt O, Shay T, Eisenstein M. Docking to single-domain and multiple-domain proteins: old and new challenges. Proteins 2005;60:195–201.
92. Murphy J, Gatchell DW, Prasad JC, Vajda S. Combination of scoring functions improves discrimination in protein-protein docking. Proteins 2003;53:840–854.