Abstract

Protein-protein complex structure prediction, or protein docking, is of wide consequence. Because of the difficulty and expense associated with experimental methods, protein docking simulations can serve not only as a test of our understanding of protein physical chemistry, but also as a tool for illuminating biological phenomena, and in engineering and biomedicine as a method for developing new protein-based technologies. To encourage the development of structure prediction algorithms, Critical Assessment of PRediction of Interactions (CAPRI), a “blind” prediction task requiring theoretical predictions before experimental verification, provides an opportunity for developers to evaluate their methods. This work reports novel methods for competing in a CAPRI subproblem, the CAPRI scoring challenge, wherein participants are provided a set of protein complexes generated by different algorithms and are tasked with the identification of correct, or near-native, structures within that set. This requires (1) a strategy for leveling the playing field for structures generated by different docking algorithms via a “relaxation” process and (2) a new score for identifying the near-native decoys from a set so processed. The level playing field allows a fair comparison between structures generated by diverse structure prediction strategies. My novel score outperforms standard scores: the average $z$-score of near-native structures is 0.59 with the novel score, compared with 0.44 and 0.43 for total and interface scores, respectively. More importantly, when the top 15 structures are selected by score, the novel score identifies, on average, 4.0 near-native decoys, compared with 2.0 and 1.5 for total and interface scores, respectively. [Word Count: 248]
1 Introduction

Almost all biological processes are mediated by protein-protein interactions. These interactions are frequently crucial to the proper functioning of the processes they mediate, and exhibit characteristic specificities that underlie their function. The structure of protein-protein complexes—like the structure of any macromolecule—can provide insight into the biological roles and mechanisms of action of the involved proteins. Most commonly, structural information comes from experimental techniques such as NMR and X-ray crystallography. Unfortunately, it is sometimes not possible to determine the structure of a protein-protein complex experimentally because such techniques typically require high protein concentrations and protein-specific tuning of experimental conditions. Even when successful, these methods are frequently time consuming and expensive. In contrast, computational methods are almost always comparatively inexpensive and less time consuming. They are also easily and more cheaply automated. If it can be reliably accurate, computational prediction of protein-protein complexes, or protein docking, can obtain this structural information without the need for experimentation. In addition, protein-protein complex structure prediction techniques could be, and have been, adapted to engineer molecules with altered or entirely novel binding characteristics.

The results of current docking simulations, however, are significantly less reliable than the results of experimental methods. They typically play the role of eliminating structures or designs that are unlikely to be correct in preparation for subsequent experiment, rather than being directly applied to the design and prediction problems. To encourage the development of methods that can be directly applied to structure prediction and design, the Critical Assessment of PRediction of Interactions (CAPRI) provides an opportunity for developers of predictive methods to test their methodology blindly.\textsuperscript{1} To ensure that participants have no knowledge of the structures they are predicting—that is, that their predictions are “blind”—CAPRI targets are always unpublished experimental structures that have been donated by generous experimentalists. Predictors are typically given the experimental unbound structure of one or both of the partners and asked to predict the structure of the bound complex. By providing an opportunity to make truly blind predictions, CAPRI allows for the investigation of the two fundamental problems of structure prediction, sampling and scoring, with a level of purity that is not usually possible.
Sampling refers to the ability of an algorithm, in this case of two proteins interacting, to quickly generate all possible protein-protein interactions that have low energy, without generating those that are not likely to have low energy. This can include the relative rigid-body orientation of the binding partners and torsion angles within each monomer and, in general, bond lengths and angles. A sampling strategy is considered good, or “efficient,” if it quickly finds many high-quality configurations, and “inefficient” if it does not. Many diverse approaches to this problem exist, from fast Fourier transform (FFT)-based geometrical complementarity matching, to second-by-second tracking of atomic velocities and accelerations due to forces in molecular dynamics (MD) simulations.

Scoring is the problem of identifying which of the configurations generated by the sampling algorithm is “best.” When making predictions about protein-protein complex structure, the best candidate structures are identified by having the lowest free energy. A scoring algorithm, or score function, is thus good if it accordingly estimates complex free energy. In many sampling algorithms, information from the score function is used to inform the direction of sampling—an example of this is would be a deterministic gradient-based optimization of torsion angles. Like sampling, successful scoring methodologies take many forms, from simply assessing shape complementarity, typical in FFT-based methods, to moment-to-moment potential energies in an MD simulation.

While the standard CAPRI challenge evaluates participating methods’ scoring and sampling, a more recently developed CAPRI challenge seeks to evaluate only participants’ scoring. This subproblem, the CAPRI scoring challenge, takes place after the standard CAPRI challenge but before the results—and the true native structure—are announced. Predictors, who only submit ten models for the prediction challenge, are invited to submit 100 of their best models for use in the scoring challenge. These structures are made anonymous and distributed to scorers. The number of prediction groups that participate usually ranges between ten and thirty, meaning that there frequently upwards of 2000 structures available for scoring, although this number varies widely from target to target. Within each set of structures, the number of correct, near-native structures (here $C_\alpha$ RMSD < 5.0 Å) also varies widely, from zero to 15%. In contrast to the standard challenge, which tests both sampling and scoring, the scoring challenge provides scorers with access to the best results of a large number of different sampling methods and allows participants to isolate problems with their score functions.
Because they are generated by a wide variety of often incompatible docking algorithms, structures distributed during the CAPRI scoring challenge are highly diverse. Past CAPRI prediction challenges have seen the use of more than 20 distinct prediction methods, including approaches as diverse as HADDOCK, which relies primarily on non-structural biochemical and biophysical information for its predictions; ATTRACT, which models residues as a few pseudoatoms and systematically enumerates rigid-body configurations; and ZDOCK/RDOCK, which together are a multiscale approach using FFT sampling followed by CHARMM MD refinement. Each participating group manipulates the docking partners in different ways. Some predictors omit any ligands, while others submit them as part of the structure. Each predictor group removes different parts of the protein that they anticipate are not involved with binding. To evaluate decoys fairly, however, they must have the same sequence, and must all either use or not use the ligand for scoring. Nevertheless, even once these issues are resolved, decoys are not on a level playing field when compared when using any particular score function, because score functions used by each group in their predictions have different characteristic preferences parameters like minimum energy distance in a van der Waals contact, bond lengths and angles, hydrogen bonds, covalent bond spring constants, and so forth. My work includes a method for first enforcing sequence-identity across all structures and then for refining, or relaxing, decoys such that an entire decoy set can be fairly compared using a single score function.

The Rosetta protein modeling package, of which Rosetta’s docking application is a part, uses an intermediate level of detail for both sampling and scoring. Proteins are modeled using a full-atom resolution representation but with idealized bond lengths and angles. Sampling does not attempt to model the time-dependent motion of atoms. Instead, the original docking application in Rosetta, RosettaDock, combines random rigid-body moves with rotamer-based side chain packing in a Metropolis Monte Carlo-plus-minimization sampling strategy that always accepts random moves that lower predicted energy and rejects changes that increase energy with probability equal to the Boltzmann-weighted probability associated with the energy change. Moves are subjected to gradient-based minimization before scoring and application of the Boltzmann-weighted probability, or Metropolis, criterion. Since the original RosettaDock was developed, sampling has continued to improve, notably by guiding backbone sampling using a conformer selection model of protein-protein binding. The conformer selection model proposes that a small ensemble of unbound monomers
exists in solution, around a minimum free energy. As a result of favorable interactions with the binding partner, those conformers with geometry compatible with binding are selected out of this ensemble. By mass action, those same conformers are then replaced in the unbound ensemble, thus “selecting” for conformers that are compatible with binding.\textsuperscript{11,12} Computationally, this is achieved by generating an conformer ensemble, and docking with one “active” member of that ensemble. The active member is randomly swapped out for one of the inactive ensemble members, and the change is accepted if it meets the Metropolis criterion.

RosettaDock’s scoring approach is based on the same full-atom representation of the docking partners used in the high-resolution phase of sampling. It explicitly scores each atom, including hydrogen atoms, but it includes statistical terms that convert the frequency of a feature seen in the Protein Data Bank (PDB) into an energy using the Boltzmann equation and Bayes’s theorem.\textsuperscript{13–16} The score, which is calculated by the score function, is the sum over a number of terms each of which models the contribution to the free energy of complex formation by particular physical phenomenon like van der Waals contacts. All score function terms are classified as either one-body (calculated based characteristics of only one residue) or two-body (based on interactions between two residues). The fact that no term calculates based on the properties of more than two interacting bodies is crucial for the speed with which the score of a structure can be calculated, which is itself critical to the efficiency of the sampling algorithm. All terms are continuous and differentiable with respect to bond torsion angles, and they are said to be “pairwise-additive,” because the total score of a structure is the simple sum over all one-body and two-body scores for all residues. The most important terms in the score function are a Lennard-Jones model for van der Waals interactions (which are broken into attractive and repulsive portions for computational convenience and independent weighting), a statistical hydrogen bond potential\textsuperscript{16} a pairwise Gaussian solvation model,\textsuperscript{17} a statistically-derived term that scores based on the expectation that two residues are found near one another ($P_{\text{pair}}$),\textsuperscript{15} and a sidechain conformation probability term based on a rotamer library.\textsuperscript{18} This score function has shown itself remarkably successful, despite these coarse-grained approximations.

Using a wide variety of differing, often custom-developed Rosetta-based sampling strategies, there have been a steady stream of docking successes emerging from Rosetta Commons labs. In the last two years alone, successful methods using Rosetta have been reported for epitope grafting.\textsuperscript{19}
novel interface design,20 peptide-protein complex structure prediction,21 and protein-biominal
derosette adsorption structure prediction and design.22 (For a general review of the entire Rosetta package’s
functionality, see Das and Baker.23) In contrast, outside of modeling non-peptide molecules (i.e.
RNA, DNA, biominals, etc.) progress on the score function has remained slow (and frequently
unpublished) owing to the myriad interrelations between score terms and most sampling strategies’
reliance on the score function for guidance. Many of the useful techniques and modern improve-
ments require the relaxation of one or more of the core requirements for a Rosetta score term, like
pairwise additivity or differentiability, and hence have not been included as a term into the score
function.24,25 Even the subtle corrections introduced by Song et al.,26 which do not involve the
addition or removal of a score term and were expressly designed not to affect sampling, have not
been widely adopted.

The two traditional, and by far the most most commonly used, versions of the Rosetta score
function for docking are total score and interface score. Total score is, as described above, the
simple sum of all score terms over all residues throughout the putative biomolecule structure (in
this case, a protein-protein complex). The implied reference state for this calculation is the unfolded
state, so the total score represents a prediction about the free energy of the transition $U \rightleftharpoons F$, where $U$ is the unfolded state and $F$ is the folded state—although it must be mentioned that
polymer configurational entropy is not explicitly modeled in Rosetta, and solvent entropy enters
only through the pairwise Gaussian solvation model, which is a coarse-grained estimate. In docking,
total score estimates the energy change of the transition is $U_A + U_B \rightleftharpoons F_{AB}$, where $A$ and $B$ are
the two polypeptide chains being docked.

Interface score, which has been shown to discriminate more effectively between native and
non-native interfaces than total score,10 uses a different, decoy-specific reference state. When
calculating an interface score, the total score of the putative complex is calculated. Then, the
monomers are separated, new sidechain rotamers are assigned to minimize energy (a process known
as “packing”),8,27,28 and the separated structures’ total score is calculated. The latter is then
subtracted from the former, yielding the energy of interaction between the structures—the so-
called “interface energy.” This represents a useful but less physically realistic transition than total
score: $F_A^* + F_B^* \rightleftharpoons F_{AB}$, where $F_A^*$ and $F_B^*$ are the separated, packed monomers and $F_{AB}$ is the
complex.
Several of the approximations made by Rosetta’s score function seem to lead to nontrivial errors. In particular, the modeling of electrostatic and solvation energies have recently come under increased scrutiny.\textsuperscript{29,30} Electrostatic interactions are currently modeled by the statistical term $P_{\text{pair}}^{14,15}$ and a distance-dependent dielectric constant electrostatics model based on Coulomb’s Law. The former does not model with atomic resolution and the latter is usually given very low weight because both charges and the dielectric constant are poorly determined. The effect of non-hydrogen bond electrostatic interactions is thus probably systematically underestimated and/or modeled with low precision in Rosetta scores. Similarly, the pairwise additivity assumption and continuum approximation made by the Gaussian solvation model\textsuperscript{17} introduces errors, particularly at interfaces, where the implicit solvent is not able to account for the water molecules frequently found there.

These problems are the result of fundamental inaccuracies, and perhaps also double-counting, in the Rosetta score function. They will not be resolved by simply tuning score function weights, or adding a new score term. Besides the obvious implementational complexity of introducing elements to the score function that are not pairwise additive or that are undifferentiable, many of the physical phenomena that one might seek to model more accurately, like the complex interactions between electrostatics and solvation, are not fully understood.\textsuperscript{31} While these problems with the Rosetta score terms will probably eventually be dealt with—by brute-force using presently computationally-intractable models (e.g. quantum-chemical simulations) on future hardware, if nothing else—we would like a way to better evaluate the near-nativeness of decoys in the mean time.

A variant of the classic score functions was recently proposed by Raveh \textit{et al.} in an investigation modeling peptide-protein interactions which combined total and interface scores in an average—in effect, summing the two scores.\textsuperscript{21} This sum score had discrimination that was better than both total and interface scores when docking peptides to proteins. This work takes a similar, but more aggressive, approach that combines total and interface score nonlinearly and that tends to emphasize interface score over total score. The result is a score that provides better enrichment of high-quality structures amongst the very best ranking decoys without lowering discrimination compared to the sum score. Both sum and elliptic scores are found to discriminate better than total and interface scores alone.
2 Materials and Methods

2.1 Parsing Decoy Format and Sequence

Outside of score function preferences, structures provided for CAPRI scoring challenges are not sequence identical and vary in their adherence to the PDB standard. To manage these issues, this work uses a set of scripts that apply ClustalW alignments to trim all sequences in a given structure set down to a user-specified minimal sequence. It also standardizes chain designations and resolves various formatting inconsistencies, such as nonstandard amino acid codes and missing atoms, typically by removing offending residues.

These protocols were implemented in Python 2.6 and require ClustalW 1.82. Code is available on the Rosetta Commons Subversion (SVN) system.

2.2 Relaxation of Decoys under Rosetta’s Force Field

The protocol used to relax input structures into the Rosetta force field has three phases, each of which is a published Rosetta 3 executable protocol in its own right. They are as follows:

1. Simulated-annealing sidechain packing, using a fast trie vs. trie algorithm, with a fixed backbone using the fixbb executable.\textsuperscript{27,28} Exact command line:

\begin{verbatim}
$PATH_TO_EXE/fixbb.static.linuxgccrelease -l $targetStructures -out:path:pdb $fixbbOutputDir -nstruct 1 -database $PATH_TO_DB -resfile $PATH_TO_RESFILE -out:file:fullatom -ex1 -ex2aro -use_input_sc -mutiple_processes_writing_to_one_directory
\end{verbatim}

2. Simultaneous rigid body moves and packing, using docking_protocol executable in high-resolution docking only mode. This executes only the high resolution search phase of the original RosettaDock rigid-body protein-protein docking algorithm. At this step, three output structures are produced for each input structure, to improve Rosetta’s chances of finding a good score minimum near the original structure.\textsuperscript{32} Exact command line:

\begin{verbatim}
$PATH_TO_EXE/docking_protocol.static.linuxgccrelease -l $fixbbOutStructures -out:path:pdb $dockingOutputDir -database $PATH_TO_DB -nstruct 3
\end{verbatim}
3. Ramp-repack-minimization cycles, with repetitive sudden drops followed by slow increases in the \textit{fa}_\text{rep} (Lennard-Jones repulsive) score weight. The default run sets the \textit{fa}_\text{rep} weight to 0.02, packs and minimizes, sets the weight to 0.25, packs and minimizes, sets the weight to 0.55, packs and minimizes, set the weight to 1.0, packs and minimizes, then starts again from the beginning. Each call to the gradient-based minimizer optimizes energy with respect to all backbone torsional angels and all rigid body degrees of freedom.\textsuperscript{33} Exact command line:

```
$PATH_TO_EXE/relax.static.linuxgccrelease -l $dockingOutStructures
-out:path:pdb $relaxOutputDir -database $PATH_TO_DB -nstruct 1 -relax:fast
-out:file:fullatom -ex1 -ex2aro -use_input_sc -relax:jump_move
-mutiple_processes_writing_to_one_directory
```

All executables were compiled using revision 42122:42133 from the Rosetta Commons SVN trunk.

2.3 Calculation of $C_\alpha$ RMSD, and Interface, Sum, and Total Scores

The reported RMSDs and both interface and total scores were calculated using default weights with the \texttt{rosetta} scripts\textsuperscript{34} executable. The calculation of scores during sampling for the refinement protocol used the default scores for that application. The sum score, adapted from Raveh\textit{ et al.},\textsuperscript{21} was calculated by summing total and interface score.

RMSD is calculated by summing the square differences in $C_\alpha$ atom position, after aligning the structures to minimize this quantity.

The weights for standard Rosetta 3 score functions—and those used for the calculation of both interface and total score—are those as used in \texttt{fa_standard}.

2.4 Evaluation of Discrimination

A standard statistical measure of a score function’s discrimination between near-native and non-native is the low-RMSD $z$-score, or $z_{\text{rms}}$. It was evaluated in the usual way:\textsuperscript{8}
\[ Z_{\text{rms}} = \frac{E_{\text{hi}} - E_{\text{low}}}{\sigma_{E_{\text{hi}}}} \]  

(1)

where \( E_{\text{hi}} \) and \( E_{\text{low}} \) are the average score of all models with high and low RMSD, respectively, and \( \sigma_{E_{\text{hi}}} \) is the score standard deviation of structures with high RMSD. For this purpose, structures have low RMSD if they are in the lowest 5% of the decoy population by RMSD and high otherwise.

Unlike Wang \textit{et al.},\(^8\) however, I excluded the highest-scoring 5% of decoys. This change was required because decoy sets from the CAPRI scoring challenges sometimes contain a handful physically unrealistic, non-native decoys with very high score (sometimes in the positive hundreds of thousands) that, if used in this analysis, make all score functions appear highly discriminating.

Another common metric of discrimination is the area under a Receiving Operator Characteristic (ROC) curve. The ROC curve plots true positive rate against false positive rate, meaning that the greater the area under the curve, the more likely that a randomly selected datum will be classified as “positive” or “negative” correctly. A “positive” structure is any structure with \( C_\alpha < 5.0 \) Å, and all others are considered “negative.” The area under the curve was computed by the Riemann sum over each point in the curve. Points in the curve were separated by 0.1 percentile units.

2.5 CAPRI Target Data

Table 1 shows the six decoy sets used in this paper. All are past CAPRI prediction experiment targets for which a scoring challenge was also performed. Past CAPRI scoring targets, such as those that included nucleic acids (targets 33, 34), dealt with more than two docking partners (target 51), for which there were no or very few correct solutions (targets 28, 30, 33, 35, 36, and 38) except target 39, for which the native structure coordinates have not been published (targets 53 and 54), or targets for which the archival scoring sets were not available (targets 28, 40, 42–49) in the scoring decoy set were not included in the analysis. Nonstandard challenges (targets 52 and 55) were also not included.

2.6 Image Software

All molecular visualization used The PyMOL Molecular Graphics System, version 1.3, a product of Schrödinger, LLC. Renderings used \texttt{ray\_trace\_mode 3} with \texttt{antialias at 2} and \texttt{cartoon\_fancy\_helices}
Table 1: Basic information for CAPRI targets used in this paper. Columns: the target number issued by the CAPRI organizers; the PDB ID for the experimental crystal structure, “native” for RMSD calculations; the highest quality structure submitted by the Gray Lab for this target during the prediction experiment; the fraction of structures in the provided data set with Cα RMSD less than 5.0 Å; the total number of structures provided by the CAPRI organizers that were not rejected during preprocessing their sequence; the fraction of structures submitted by scorers that were acceptable (*) or better; a brief description of the target complex’s constituents.

<table>
<thead>
<tr>
<th>Target Number</th>
<th>PDB ID</th>
<th>Gray Perf.</th>
<th>% &lt; 5.0 Å RMSD</th>
<th>Ntot</th>
<th>Scorer Perf.</th>
<th>Protein Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>2VDU (BE)</td>
<td>-</td>
<td>8.1%</td>
<td>1481</td>
<td>19.7%</td>
<td>tRNA M7G methylation complex</td>
</tr>
<tr>
<td>32</td>
<td>3BX1 (BD)</td>
<td>***</td>
<td>1.8%</td>
<td>593</td>
<td>0.017%</td>
<td>subtilisin/subtilisin inhibitor complex</td>
</tr>
<tr>
<td>37</td>
<td>2W83</td>
<td>-</td>
<td>3.0%</td>
<td>1404</td>
<td>26.3%</td>
<td>Arf6 GTPase/effector Jip4</td>
</tr>
<tr>
<td>39</td>
<td>3FM8</td>
<td>-</td>
<td>0.16%</td>
<td>1201</td>
<td>44.3%</td>
<td>centaurin alpha-1/FHA (domain KIF13B)</td>
</tr>
<tr>
<td>41</td>
<td>2WPT</td>
<td>**</td>
<td>13.8%</td>
<td>567</td>
<td>53.1%</td>
<td>IM2/colicin E9 DNase</td>
</tr>
<tr>
<td>50</td>
<td>3R2X</td>
<td>-</td>
<td>8.4%</td>
<td>1325</td>
<td>25.0%</td>
<td>binding protein HB36.3/1918 influenza HA</td>
</tr>
</tbody>
</table>

on. The cartoon_sampling parameter was set to 10.

All plots were generated with the R language for statistical computing, by the R foundation. Of particular interest are the rainbow() function, which determined point coloring in Figures 2 and 3. Image editing after production with R or PyMOL used the GNU Image Manipulation Program (GIMP), available at http://www.gimp.org/.

3 Results

3.1 A Three Step Relaxation Protocol Eliminates Most Decoy Shelving and Lowers Score Spread

Once all structures are sequence-identical and adhere to the pdb standard, structures must be scored. Figure 1, left, shows the results of applying the Rosetta all-atom force field to target 29. The total score versus Cα RMSD plot shows at least four, nearly discrete score tiers: -670 to -650; -650 to -630, -625 to -595, and -590 to -570. The higher three of these four score tiers contain structures that are near-native, but are not low-ranking. In a blind prediction experiment, these good solutions to the CAPRI scoring challenge would be hidden by the lower tier.

Because of the tiers’ wide, flat appearance on score versus RMSD plots, this phenomenon is termed “shelving.” Shelving was eliminated by developing a protocol that relaxes structures into Rosetta’s unique all-atom force field such that shelving is eliminated by combining three preexisting protocols.
Figure 1: Plots showing extent of shelving with (right) and without (left) the final relaxation protocol, using CAPRI target 29 (see Table 1). Left, using simulated annealing with the rotamers-trials packer, producing 3 structures for each input structure (fixbb protocol). Right, the final relaxation protocol outlined in Section 2.2. In both the left and center plots, several distinct score shelves, which appear to be independent of C\textsubscript{α} RMSD, are easily visible.

Protein docking simulations in Rosetta usually require relaxed input structures, unless those structures were generated by Rosetta itself. When the input structure is derived from x-ray crystallography data, for example, relaxation takes the form of “prepacking,” which consists of the high-resolution phase of the standard RosettaDock docking algorithm.\textsuperscript{9} For this application, however, prepacking of that kind was insufficient to remove the shelving.

Through experimentation with numerous different combinations of Rosetta-based local sampling techniques (i.e., those that sample only very near the input structure), a protocol that provides empirically satisfactory results—that is, shelving has been eliminated from the score versus C\textsubscript{α} RMSD plot (Figure 1, right)—was eventually reached. The final protocol consists of thorough repacking, followed by simultaneous repack and rigid body refinement, followed by ramped repack-minimization cycles (Section 2.2). Other techniques examined before settling on the protocol in Section 2.2 were single calls to a gradient-based minimizer acting simultaneously on torsions angle and rigid-body displacement, prepacking followed by the same, and prepacking followed by packing and rigid-body Monte Carlo moves. The most important and productive step in the protocol is
probably the third step, which contains ramp-repack minimize cycles.

3.2 Near-native Structures Have Low Interface and Total Energy, Leading to the Definition of an Elliptic Combination Score

Figure 2, columns one and two, show plots of total and interface scores, respectively, versus $C_\alpha$ RMSD. When RosettaDock has converged, there is typically a dense band of structures deemed incorrect across the center of the plot, and one or more descending funnel-shaped regions that descends below the band of incorrect structures towards low score and, when Rosetta is correct, low RMSD. Because of its tapered, downward-pointing appearance, such a structure is termed a “funnel.” Figure 2, row two, column one, has a fairly usual, correct funnel. Figure 2, column two, row two, also has several funnels, but they are not correct. In a blind prediction scenario, because RMSD information is not available, it is difficult to discern which funnels are correct and which are not.

A plot that is available in a blind prediction scenario, however, is shown in Figure 2, rightmost column. These are plots of interface score versus total score. Because this is not a blind study, each point is colored by its $C_\alpha$ RMSD rank, with the lowest decoy colored red, and the highest decoy colored purple; intermediate decoys are colored in using the ordering of the spectrum of light. For most targets, the lower left-hand corner of these plots is enriched significantly for near native decoys.

To preferentially select decoys based on a criterion of closeness to the “lower left-hand corner,” a quantitative way of comparing closeness to the lower left hand corner is required. The sum score, $S_{\text{sum}}$, which is computed using $S_{\text{sum}} = S_{\text{interface}} + S_{\text{total}}$, does so by assigning the $L_1^{\text{norm}}$, or Manhattan distance from the origin in the interface score/total score plane, as a score to each point. This function produces level sets that are lines with negative slope in the third quadrant of that plane. I propose a different score which has elliptical, rather than linear, level sets that are not centered around the origin.

Most intuitively, the elliptic score represents a distance in a transformed interface score/total score plane (Figure 3). The transformation converts the elliptic level sets into circular level sets, which allow visualization of the score as a Euclidean distance. The transformed plane is generated by scaling the interface score direction by the ratio of the total and interface scores’ interquartile
Figure 2: Each row shows various plots for each of six refined, previous CAPRI scoring targets. Top to bottom, the targets are 29, 32, 37, 39, 41, and 50. Each column shows one of five plots of interest. Left to right, the plots are total, interface, sum, and elliptic scores (in Rosetta energy units) versus $C_\alpha$ RMSD (in Å), and a plot of interface score versus total score.
ranges such that the transformed interface score $\hat{I}$ is computed by the formula

$$\hat{I} = \frac{IQR_T}{IQR_I}$$

(2)

where $IQR$ is the interquartile range ($Q_3 - Q_1$) for the total and interface scores, as indicated by subscripts $T$ and $I$, respectively.

The transformed interface score/total score plane is shown in Figure 3. The point $p$ is the lowest total ($T_{\text{min}}$) and transformed interface scores ($\hat{I}_{\text{min}}$) amongst all decoys, or the point ($\hat{I}_{\text{min}}, T_{\text{min}}$). The distance between this point and a decoy’s position in the total score/scaled interface score plane is the elliptic score, $E(I, T)$:

$$E(I, T) = \sqrt{(T - T_{\text{min}})^2 + \left( (I - I_{\text{min}}) \times \frac{IQR_T}{IQR_I} \right)^2}$$

(3)

where $T$ and $I$ are the total and interface scores of the decoy, respectively. As an example, two points, $p'$ and $p''$ are shown on Figure 3 with their respective distances from $p$, $r'$ and $r''$. These distances are equal to the elliptic score assigned to these decoys. Elliptic score level sets, each
separated by 10 points, are indicated as black circles on this plot. Because $E$ has concentric, elliptically shaped level sets in the total score/unscaled interface score plane, $E$ is referred to as the “elliptic combination score.”

### 3.3 The Elliptic Score Makes Improvements Over Sum, Interface and Total Scores

Figure 2, columns 1–4, show plots of score versus $C_{\alpha}$ RMSD for total, interface, sum, and elliptic scores, respectively. Once decoys are relaxed into Rosetta’s force field, correct funnels are not necessarily present, and the scoring is quite noisy, particularly for total and interface score individually. Both scores on most targets show at least a few stray decoys that, outside of a funnel and above 5 Å $C_{\alpha}$ RMSD, have very low energies. Many show non-native funnels (target 50, total score; target 39, interface score), while others fail to show any evidence of a native funnel whatsoever (target 41, total score).

Comparing the elliptic score to total score, there is qualitative improvement in two major capacities: the removal of a few non-native funnels and the lowering of native funnels. Other qualitative improvements include the elimination of the trend toward lower energy with higher $C_{\alpha}$ RMSD in target 37, the lowering of the native funnel in targets 29 and 37, and the removal of several false-positive funnels (around 22.5 Å in target 29 and 22 Å in target 50). Clear native funnels in total score/$C_{\alpha}$ RMSD space (i.e., targets 29, 32, and possibly 37) are preserved in elliptic score/$C_{\alpha}$ RMSD space.

Compared to the interface score (Figure 2, column 2), the elliptic score has much fewer noisy, outlying false positive decoys—that is, structures that have very good total scores but that are far from the native funnel and not part of a false positive funnel themselves. In addition, many non-native funnels from interface score/$C_{\alpha}$ RMSD space are not present in the elliptic score—of particular note is the relative diminution of the multi-funnel region between 14 and 23 Å $C_{\alpha}$ RMSD in target 32 interface score. Positive outliers, however, can also be removed by this procedure (for example, the target 50 decoy at (2.5 Å RMSD, $-45$ REU).

The sum score (Figure 2, column 3) does not qualitatively differ from total score in the shape of the decoy clouds for any of the targets, except insofar that some of the noise present in total score is cleared from sum score. Unlike elliptic score, it retains the trend toward low score with
3.4 Elliptic Score Improvements Diagnose Score Function Error

One of the major improvements that elliptic score makes over interface score on target 32 (Figure 2, column two, row two) is the removal of a non-native funnel. The non-native funnel in question appears at about 16 Å Cα RMSD. The incorrect structure (Figure 4, right) uses the same binding site as the native structure (Figure 4, left), but the binding partners are rotated about 180° along the line between their centers of mass. While the predicted binding energy still comes primarily from the insertion of a loop into a binding pocket, the incorrect structure has additional, distal contacts between the surface of one partner and the N-terminus of the other. This incurs a large intramolecular penalty in total score, but is not captured in interface score.

On the other hand, one of the illustrative failures of total score was on target 50 (Figure 2, column one, row six), where there is a non-native funnel present around 21 Å. In this case, the internal monomer energy of a better-refined hemagglutinin overpowered the signal from the interface. The non-native, total score-favored interface is shown in Figure 5, and it included a buried arginine side chain with unsatisfied hydrogen bond donors, which are very unusual in the hydrophobic core and hence unlikely at a hydrophobic interface like that present in the target 50 native conformation. While this arginine residue is adjacent to an aspartate residue that may mitigate this effect somewhat, the residues are near each other both bound and unbound, by virtue of being on the same monomer. Thus, the burial of these residues at the interface likely results in a net penalty to binding. In addition, the total score-pathological interface had an exposed tyrosine and tryptophan residues (which are buried in the native conformation) and buried a lysine residue. Despite these flaws, however, the total score was sufficiently lowered by increased atomic van der Waals contacts throughout the protein that the poor quality of the interface was not recognized by total score. This was also the case for the sum score. Interface score, which is not sensitive to intramolecular changes of that kind, did not have the same issues, and correctly scored the pathological interface more than 10 points (about 40%) worse than the nearest-native interface. As a result, the elliptic score does not have that non-native funnel.
Figure 4: A rendering of the docking partners in target 32 in the correct (left) and interface pathological (right) configurations. Note the circle, indicating the extra, incorrect N-terminus contacts in the interface pathological configuration.

Figure 5: A rendering of the docking partners in target 50 in an incorrect, target score-pathological state. Hemagglutinin is colored light green, the designed binder is colored light blue. Exposed hydrophobic residues are indicated in red, buried ionizable residues are indicated in purple, and predicted hydrogen bonds are shown as black dashed lines.
Figure 6: A plot of the enrichment of near-native decoys (those with Cα RMSD < 5 Å) versus fraction of decoys used for selection—hence, the point at 0.1 represents the enrichment of the lowest 10% of scorers for each target used in this study. Red trace is total score, green is interface score, cyan is sum score, and blue is elliptic score. Note that maximal value on the y-axis for target 32 is almost an order of magnitude larger than for all other targets.

3.4.1 Elliptic Score Outperforms Total and Interface Scores When Score-Based Selection of Correct Decoys Is Required

Figure 6 shows the enrichment of near native (Cα RMSD < 5.0) by the fraction of decoys selected. Unsurprisingly, either interface score or total score fails to enrich significantly on most targets, and all scores fail on target 39. Both total and interface scores actually show enrichment less than one in at least one case. Where total score is successful, the sum score is somewhat more successful. Where the total score is unsuccessful, however, the sum score is typically also unsuccessful. Except with target 39, elliptic score never fails to provide an enrichment of twofold or more in the first few tenths of a percentile selected.
I calculated the discrimination of all three scores for all targets and compared them. Figure 7 shows the relationship between the z-score of the elliptic score and the z-scores of each other score for all targets. The data show that, in the average case, the elliptic score has a better z-score than each other score, although this effect is marginal when comparing elliptic and sum score (Table 2).
ROC AUCs (Section 2.4), another traditional measure of discrimination, were also calculated for all four scores on all targets (Table 2). Once again, in the average case, elliptic score is better than total and interface score, although marginally so. It is slightly worse than sum score, but it is more consistent (with a twofold lower standard deviation).

3.5 Blind CAPRI Scoring Results Using Similar but Unstandardized and Un-refined Approaches Are Not Unsuccessful

Approaches very similar in spirit, but different in application, were used to participate in CAPRI scoring rounds 24 and 26 on targets 50, 51, 53 and 54—that is, predictions were submitted blindly. On targets 50 and 51 scoring predictions were made using visual inspection guided by the total score/interface score plot (Figure 2, righthand column). For target 50 one medium and three acceptable structures were submitted. For target 51, one acceptable structure was submitted. On target 53, using an IQR-weighted combination of total and interface scores, one medium scoring prediction was achieved. No correct structures were submitted by scorers for target 53, and only 19 of 1400 structures distributed for scoring were acceptable or better.

With the exception of target 51, the relaxation protocol described in section 2.2 was used for all targets. Target 51 was processed differently because is a three domain protein. The experimental coordinates for targets 51, 53, and 54 are not published at the time of this writing, and so were not used in the other analyses presented throughout this paper.

4 Discussion

This paper presents solutions to several problems addressing the CAPRI scoring challenge. Errors in form and differences of protein sequence were removed by developing a set of scripts that automate the process of handling inhomogeneous input structures. Inhomogeneity resulting from differing protein models was removed using an aggressive, three-step protocol that sufficiently relaxed input structures under the Rosetta force field to eliminate shelving. While shelving is neither a theoretically rigorous nor a quantitative estimate of the “fairness” of a relaxation process, it will suffice for this application because it is not known which algorithms generated which structures and it is not possible to know that all decoys are being fairly compared. Given the large drop (an
improvement, since the score is meant to correspond loosely to energy) total score experienced by all decoys, however, it is not likely that much of the initial, Rosetta-incompatible character—which created the shelving to begin with—remains in the decoys. The protocol thus seems empirically, if not theoretically, satisfactory.

The aggressive relaxation required to achieve the elimination of shelving generated significant noise and non-native funnels (Figure 2, columns one and two) in the data. In particular, the minimization of backbone torsion angles during the relax protocol is probably both what eliminates shelving and introduces the noise. This effect is one of the major drawbacks to the refinement protocol, and it would be beneficial to develop a refinement protocol that refines less aggressively but retains the ability to eliminate shelving.

Pathologies introduced by aggressive relaxation, however, inspired the definition of a new score that capitalizes on the advantages of both total and interface score—namely that total score is high when the backbone configuration is unrealistic and the interface score is high when the interface is unrealistic. The elliptic score abandons the usual Rosetta analogy to energy and instead considers the information present in total and interface scores.

The total score reports on the physical realism of the entire structure: how likely is it that this complex, including the intramolecular forces in each monomer, is physical? Because the total score is an extensive function, the signal from the interface can be drowned out by the signal from non-interfacial regions. This occurred in target 50, where the lowered energy of the entire hemagglutinin structure biased total score toward a non-native funnel. Interface score was insensitive to these pathological changes, and it resulted in the removal of this non-native funnel in the elliptic score.

The interface energy, by contrast, evaluates only the interaction: assuming the given conformer is the binding conformer, how likely is this interface to form? Because the backbone may move during binding, this reference state is not necessarily a good measure of binding energy—the more the backbone moves, the worse the interface score performs. Target 32 exemplified this effect, where the interface score gave an unphysically low score to a structure with contacts involving an unfolded N-terminus, which compensated for a less-well packed binding pocket. Total score, on the other hand, was not sensitive to solvation penalty of unfolding the N-terminus and gave the structure an accordingly high score. The elliptic score captured this total score penalty and did not predict that this false-positive structure was low energy.
For a target protein well modeled by conformer selection, both total and interface score must be low. The total score must be low to ensure that the binding conformer is sampled with sufficient frequency in the unbound state, and the interface score must be low to ensure that the complex forms. From a strict statistical mechanics standpoint, however, the sum of the energies (as estimated by total score) should be sufficient to capture both propensities. The hypothesis that drives the elliptic score is that in the Rosetta score function, the balance between the energy associated with the protein core and the energy associated with the protein surface is not always even. By focusing on the information content of interface and total scores, the elliptic score represents an attempt to sidestep the question of how to “fix” solvation and electrostatic scores.

Recently, Raveh et al. used a similar approach that combined interface and total scores linearly and with equal weight: \( S_{\text{sum}} = \frac{1}{3}(T + I + P) \), where \( P \) is an estimate of the energy of the peptide—this value does not exist in the general docking problem, and hence cannot be applied outside peptide-protein interactions except insofar that it is already part of total score. Because scores are comparative and Rosetta Energy Units are not one-to-one with any physical unit of energy, the factor of one third is not meaningful and will be omitted. Without the peptide score, combining the scores in this way is equivalent to estimating the free energy difference in the transition

\[
U_P + U_A + F_P^* + F_A^* = 2F_{PA},
\]

or,

\[
\frac{1}{2} (F_A^* + U_A) + \frac{1}{2} (F_P^* + U_P) \rightarrow F_{PA}
\]

where \( P \) is the peptide that \( A \) binds (this energy enters through total score, not through peptide score). The latter transition begins with \( P \) and \( A \) in a state that is half folded state and half interface score reference state. This thermodynamic model of the binding transition is not physically plausible, because there must be a large energy gap between folded and unfolded states.\(^{45}\) In spite of this, however, it was more discriminating than either total or interface score individually, both when applied to peptide-protein systems\(^{21}\) and to protein-protein systems, as demonstrated in this paper.

Taken out of a thermodynamic model, and considered instead in terms of the information provided by the constituent scores, the equal-weight approach used by Raveh et al. considers the total and interface scores to contain information in proportion to the magnitude of their variability.
Because total score is typically somewhat more variable than interface score, modulating total score by interface score is a conservative choice, given that total score is probably more physically realistic.

The elliptic score is a slightly different, more aggressive, attempt to rebalance the weighting between the two scores based on their information content, which is taken to be proportional to the variability in the overall distribution of those two scores. To do this, the scores must be weighted such that they contribute more or less equally to the final score and this is achieved by weighting the scores by their interquartile ranges. In addition, the nonlinear summation of the two constituent terms in the elliptic score leads to a stiff penalty for structures that are low energy by only one of the constituent scores, because the elliptical level sets arc back away from the extrema of the decoy cloud in the total score/interface score plane (Figure 3).

The elliptic score is the first score that incorporates information from total and interface score that is formulated specifically for the general docking problem. This equal weighting of information improves the fraction of decoys that are correct in the lowest score quantiles—a property that is very important for selecting structures for experimental characterization or submission in a structure prediction challenge—while improving overall discrimination.

This method may not be the best method of combining interface and total score but it demonstrates that such a combination, especially a nonlinear combination, can be productive. There are several ways in which future work to improve on what has been presented here. Because it is always positive and does not depend linearly on the various score terms that the underlying energy functions use (e.g. fa.atr), which are meant to capture real physical phenomena, the elliptic score is not analogous to energy, even in the limited way that Rosetta Energy Units are. Historically, however, this has not been of overwhelming concern. Only total score retains a truly well-formed analogy to energy, as both interface and sum scores use unrealistic, decoy-specific reference states.

In contexts where the analogy to energy is not important, there may still be better ways of modulating the total and interface scores to appropriately combine them. A linear combination between total and interface scores—effectively a weighted sum score—was also tested and found not to be as discriminating as the elliptic combination (an IQR-weighted $L^2_{\text{norm}}$). Perhaps the difference in efficacy arises from the reduced importance of the transition $F^*_p \equiv F_p$ in the peptide model, or the nonlinear sum used to calculate elliptic score. Perhaps higher degree $L^p_{\text{norm}}$ functions would
be more effective, like the maximum function ($L_\infty$ norm). Additionally, weighting by the ratio of the two scores’ IQRs may not provide the best estimate of the relative value in the elliptic score. Other functions for normalizing the two scores to contribute equally, like standard deviation or variance may prove to be more useful, as may incorporating knowledge about the relative discriminatory powers of the two score functions—that is, more explicitly weighting the two scores unequally, or with a dependence on a problem-specific parameter(s). For example, it seems likely, based on the observations made about target 50, that total score is less accurate for larger protein size to interface size ratios.

In the future, this approach must be tested on more numerous and varied docking problems and, most importantly, blindly in the context of a CAPRI scoring experiment. Given, however, that even visual inspection-based approaches based on interface score and total score combinations are able to perform reasonably well in a blind CAPRI scoring challenge, these results are very promising. Nevertheless, the data presented here represent a small subset of the interesting problems to which Rosetta can be applied, even within the realm of docking (no antibodies, for example, were tested). As a result, the results presented here may not generalize to all protein prediction problems. Instead, this paper presents the elliptic score as a promising proof-of-principle, and provides a hypothesis to explain these preliminary results.

The elliptic score was developed primarily to select a very small number of highly accurate decoys from a large pool for blind scoring challenges. The systems developed in this paper allow for the rapid, highly-automated selection of a small, well-enriched subset of structures from a large set with few correct solutions. This system solves many of the problems that have made participation in the CAPRI scoring challenge using Rosetta very difficult in the past.

The elliptic score, however, has applications that are significantly more broad. The task to which the elliptic score is best suited—the selection of a very small number of high quality structures—appears frequently in practice. Protein designers in particular always must select a small subset of their designs to validate experimentally, and even when a directed evolution or library screening approach is being used, there are almost always too many designs to test. To make selections, protein designers frequently use information from total and interface and total scores, but do not usually combine them quantitatively. The elliptic score defined here provides a tested method for doing so that may prove helpful for protein design as well as blind structure prediction. While the
elliptic score is not a satisfactory substitute for a proper, physically-inspired score function, until the model inaccuracies and score double counting in the Rosetta score function are resolved, this approach and others like it can stand in for a score function that estimates true free energies.

5 Acknowledgments

I thank the scientists and programmers who have developed the Rosetta software prediction suite since its inception. Members of the Jeffrey Gray group at Johns Hopkins University and William Schief group at the University of Washington were helpful not just for their contributions to the Rosetta suite but also for insightful conversations and guidance throughout my becoming a part of the Rosetta community. Particularly crucial were Brian Weitzner, Aroop Sircar, Bruno Correia, Monica Berrondo, and Sergey Lyskov for their knowledge and helpful advice regarding navigating and using Rosetta and Sidhartha Chaudhury, Mihai Azoitei, Mike Pacella, and Krishna Kilamibi for stimulating conversations regarding biophysics more generally, and Professors Jeffrey Gray and William Schief for their mentorship and scientific wisdom.

I would also like to thank Hannah Joo for her invaluable editorial assistance and insightful discussions throughout the process and Professors Juliette Lecomte and Karen Fleming for their advisory roles in my education more generally, and gratefully acknowledge the support of the International AIDS Vaccine Initiative in the Schief laboratory and the high-performance computing infrastructure provided for by National Institutes of Health grant R01-GM078221.

Finally, I would like to thank my parents Kathy Perko and Trevor Porter for their guidance, encouragement, and support for the last 23 years.
References


