Drug Design

Comparison of Docking and Scoring Methods

• A widely spread concept is that the major weakness of today's docking programs lies not in sampling methods but in scoring functions.

• Considerable efforts have been devoted to the development of computational methods for describing protein-ligand interactions.

• The different approaches can be roughly grouped into three categories:

- force field methods
- empirical scoring functions
- and knowledge-based potentials.

Comparison of Docking and Scoring Methods

- All of these scoring functions have been validated on various sets of protein-ligand complex structures.

- Actually, several comparative studies of various scoring functions have already been published.

- These studies certainly represent one valid approach for evaluating scoring functions in a molecular docking context.

- But a potential drawback in these studies is that they emphasize more on the overall performance of a complicated procedure in which the docking algorithm and the scoring function are coupled together.

Comparison of Docking and Scoring Methods

- If a certain docking/scoring combination fails, it is not always clear which one should be blamed: the docking algorithm, the scoring function, or both.

- Therefore, scoring functions themselves are not fairly compared in this way.

Comparison of Docking and Scoring Methods

Wang et al., J. Med. Chem.,2003

- The objective of this study is to conduct a fair evaluation of various scoring functions in the context of molecular docking.

- The central idea is to isolate the conformational sampling procedure from the scoring procedure so that all of the scoring functions can be compared on the same ground.

- To achieve this, an ensemble of docked conformations of each ligand molecule is generated by using the AutoDock program.

- Considerable efforts are made to ensure that this conformational ensemble achieves diversity rather than focuses on a few energy minima.

- Then, all of the scoring functions under test are applied to score the conformational ensemble.

Comparison of Docking and Scoring Methods

- 11 popular scoring functions on a wide spectrum of 100 protein-ligand complexes have been tested.

- The performance of each scoring function is evaluated by how well it reproduces the experimentally determined structures and binding affinities.

- The strength and weakness of these scoring functions are discussed.

- Consensus scoring, as a practical strategy for improving docking accuracy, is also explored.

Comparison of Docking and Scoring Methods

Preparation of the test set:

- The test set used in this study is constructed from 230 protein-ligand complexes.

- All of these complexes have crystal structures and experimentally measured Ki or Kd values.

- Only complex structures with resolution better than 2.5 Å are considered, which are 172 in total.

- Each complex is then subjected to an exhaustive conformational sampling procedure.

- One hundred complexes have passed this procedure and are included in the final test set.

Comparison of Docking and Scoring Methods

Preparation of the test set:

- Forty-three different types of proteins are presented in this test set.
- Molecular weights of ligand molecules range from 122 to 913.
- Numbers of rotatable single bonds in ligand molecules range from 0 to 20.
- Dissociation constants of these complexes range from 1.49 to 10.15 (in -log Kd or -log Ki units), spanning nearly nine orders of magnitudes.
- All ligand molecules bind to their target proteins noncovalently.
- Coordinates of all the complexes are downloaded from the Protein Data Bank.

Comparison of Docking and Scoring Methods

Conformational Sampling Procedure:

- The AutoDock program is employed to generate an ensemble of docked conformations for each ligand molecule.

- This program uses a genetic algorithm (GA) for conformational sampling.

- Each GA run outputs a single docked conformation as the final result.

- Since a conformational ensemble is desired, 100 individual GA runs are performed to generate 100 docked conformations for each ligand.

Comparison of Docking and Scoring Methods

Conformational Sampling Procedure:

• Since this conformational ensemble forms the basis for all subsequent scoring function evaluations, we expect it to depict the conformational space of the ligand (with respect to the protein) as completely as possible rather than focus on a few energy minim that are particularly favoured by AutoDock.

Comparison of Docking and Scoring Methods

Conformational Sampling Procedure:

• To achieve this goal, the following criteria have been applied to monitor the quality of the final conformation ensemble generated by AutoDock:

❑ (i) Root-mean-square deviation (rmsd) values (calculated by using the experimentally observed bound conformation as the reference) of all the docked conformations should spread throughout a wide range, e.g., 0-15 Å.

❑ (ii) The number of distinctive conformational clusters (counted by AutoDock using a clustering criterion of 2.0 Å) should fall between 30 and 70. This further ensures the diversity of the ensemble.

❑ (iii) A number of conformations should be close enough to the experimentally observed conformation (rmsd <= 2.0 Å). This ensures a proper sampling of the global minimum

Comparison of Docking and Scoring Methods

Conformational Sampling Procedure:

For some complexes, a satisfactory conformational ensemble is not obtained even at this level of computation.

Typically, the ligands in these cases are large flexible molecules, such as oligopeptides, and therefore may need even more extensive conformational sampling.

These complexes, 72 in total, are not included in our final test set.

Comparison of Docking and Scoring Methods

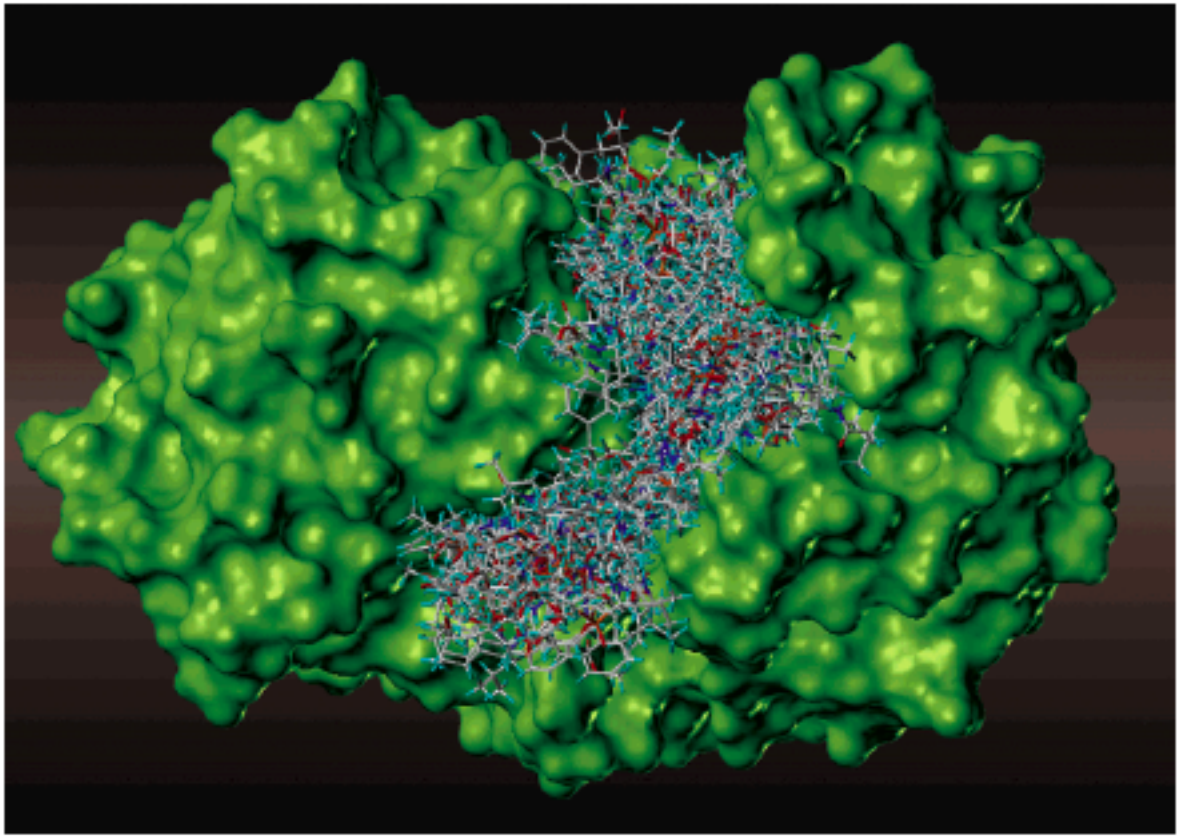
Conformational Sampling Procedure:

For all of the successful ones, 100 in total,
we then add the experimentally observed bound conformation
of the ligand to the 100 AutoDock generated docked conformations.

This further ensures the completeness of the conformational
ensemble because AutoDock may not have generated
exactly the same conformation.

This conformation should not be missed because it represents the true global minimum and is probably the most important spot on the energy surface.

Comparison of Docking and Scoring Methods



Conformational ensemble of the ligand molecule generated by AutoDock (PDB entry 1BXO).

Penicillopepsin catalyses transpeptidation reactions

Conformational Sampling Procedure:

The total number of docked conformations of each ligand thus becomes 101. These conformations usually cover the entire binding pocket and its vicinity area.

Comparison of Docking and Scoring Methods

Scoring Procedure:

- Eleven scoring functions have been tested, including the scoring function implemented in the AutoDock program.

- They can be roughly grouped into three categories:

  - (i) force field based methods, i.e., AutoDock, G-Score and D-Score

  - (ii) empirical scoring functions, i.e., LigScore, PLP, LUDI, F-Score, Chem-Score, and X-Score

  - (iii) knowledge-based potentials of mean force, i.e., PMF and DrugScore.

Comparison of Docking and Scoring Methods

Scoring functions: force-field

(1) AutoDock. The overall docking energy of a given ligand molecule is expressed as the sum of intermolecular interactions between the complex and the internal steric energy of the ligand.

$$\Delta G =$$
$$C_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)$$
$$+C_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}}\right)$$
$$+C_{elec}\sum_{i,j}\frac{q_i q_j}{\varepsilon(r_{ij})r_{ij}}$$
$$+C_{tor}N_{tor}$$
$$+C_{sol}\sum_{i,j}(S_i V_j+S_j V_i)e^{(-r_{ij}^2/2\sigma^2)}$$

- Lennard-Jones 12-6 dispersion/ repulsion term
- Directional 12-10 hydrogen bonding term
- Screened electrostatic potential
- Loss of degrees of freedom upon binding
- Desolvation term

Comparison of Docking and Scoring Methods

Scoring functions: knowledge-based

• A set of distance-dependent interaction potentials for various atom pairs.

• Only the interactions which are observed with a high frequency are considered as favourable.

• Both enthalpic and entropic effects are assumed to be included implicitly in this potential.

• The protein-ligand interaction energy is then defined as a sum of potentials over all heavy atom pairs between the complex:

$$\Delta W_{AB}(R_C) = -RTln[p_{AB}(r \leq R_C)/p_{XX}(r \leq R_C)]$$

Comparison of Docking and Scoring Methods

Scoring functions: empirical

$$\Delta G_{\text{bind}} = \Delta G_{\text{H-bond}} \sum_{\text{H-bond}} f(\Delta R, \Delta\alpha) +$$
$$\Delta G_{\text{ionic}} \sum_{\text{ionic}} f(\Delta R, \Delta\alpha) +$$
$$\Delta G_{\text{hydrophobic}} \sum_{\text{hydrophobic}} |A_{\text{hydrophobic}}| +$$
$$\Delta G_{\text{rotor}} N_{\text{rotor}} + \Delta G_0$$

Comparison of Docking and Scoring Methods

Docking Accuracy:

• The most straightforward method for evaluating a scoring function in terms of docking accuracy is to inspect how closely the best-scored (or the lowest-energy) docked conformation predicted by this scoring function resembles the one observed in the experimental complex structure.

• Here, a prediction is successful if the rmsd value of the best scored conformation is less than or equal to 2.0 Å from the experimentally observed conformation.

Comparison of Docking and Scoring Methods

Docking Accuracy:

- Success rates of all 11 scoring functions are listed in the Table.

- If using the AutoDock scoring function (success rate =62%) as reference, one can see that six scoring functions, i.e., PLP, F-Score, LigScore, DrugScore, LUDI, and X-Score, give better results (success rates ranging from 66% to 76%) while the other four scoring functions, i.e., PMF, G-Score, ChemScore, and D-Score, do not (success rates ranging from 26 to 52%)

Table 2. Success Rates of 11 Scoring Functions under Different rmsd Criteria

	success rate (%)				
scoring function[a]	rmsd ≤1.0 Å	rmsd ≤1.5 Å	rmsd ≤ 2.0 Å	rmsd ≤2.5 Å	rmsd ≤3.0 Å
Cerius2/PLP	63	69	76	79	80
SYBYL/F-Score	56	66	74	77	77
Cerius2/LigScore	64	68	74	75	76
DrugScore	63	68	72	74	74
Cerius2/LUDI	43	55	67	67	67
X-Score	37	54	66	72	74
AutoDock	34	52	62	68	72
Cerius2/PMF	40	46	52	54	57
SYBYL/G-Score	24	32	42	49	56
SYBYL/ChemScore	12	26	35	37	40
SYBYL/D-Score	8	16	26	30	41

[a] Scoring functions are ranked by their success rates at rmsd ≤ 2.0 Å.

Comparison of Docking and Scoring Methods

Docking Accuracy:

- Success rates of all 11 scoring functions under other rmsd criteria (1.0-3.0 Å) are also listed.
- It is not surprising that the success rates of all the scoring functions drop under a tighter criterion and increase under a looser criterion.
- However, rankings of these scoring functions generally do not change during this process.

Table 2. Success Rates of 11 Scoring Functions under Different rmsd Criteria

	success rate (%)				
scoring function[a]	rmsd ≤1.0 Å	rmsd ≤1.5 Å	rmsd ≤ 2.0 Å	rmsd ≤2.5 Å	rmsd ≤3.0 Å
Cerius2/PLP	63	69	76	79	80
SYBYL/F-Score	56	66	74	77	77
Cerius2/LigScore	64	68	74	75	76
DrugScore	63	68	72	74	74
Cerius2/LUDI	43	55	67	67	67
X-Score	37	54	66	72	74
AutoDock	34	52	62	68	72
Cerius2/PMF	40	46	52	54	57
SYBYL/G-Score	24	32	42	49	56
SYBYL/ChemScore	12	26	35	37	40
SYBYL/D-Score	8	16	26	30	41

[a] Scoring functions are ranked by their success rates at rmsd ≤ 2.0 Å.

- Notably, PLP, F-Score, LigScore, and DrugScore perform the best in this test. Their success rates are all above 70% with rmsd <= 2.0 Å and can still stay above 50% even with rmsd <= 1.0 Å.
- Considering the remarkable diversity presented in the test set, the performance of these scoring functions is very impressive.

Comparison of Docking and Scoring Methods

Docking Accuracy:

- One may also want to examine other conformations rather than only the best-scored one.
- The success rates of almost all of the scoring functions will improve considerably if the second or even the third best-scored conformation is taken into account.
- This can also be interpreted as that when the true conformation is missed as the very best one in binding score, it will probably appear as the second or the third best one.

Table 3. Success Rates of 11 Scoring Functions When Considering Multiple Conformations

	success rate (%) when considering		
scoring function[a]	only the best conformation	the best two conformations	the best three conformations
Cerius2/PLP	76	87	88
SYBYL/F-Score	74	89	90
Cerius2/LigScore	74	78	82
DrugScore	72	82	86
Cerius2/LUDI	67	80	85
X-Score	66	78	79
AutoDock	62	74	78
Cerius2/PMF	52	59	64
SYBYL/G-Score	42	58	66
SYBYL/ChemScore	35	47	51
SYBYL/D-Score	26	45	56

[a] Scoring functions are ranked by their success rates when only the best-scored conformation of each ligand is considered.

Comparison of Docking and Scoring Methods

Docking Accuracy:

• If considering the best three conformations in each case, five scoring functions, i.e., PLP, F-Score, LigScore, DrugScore, and LUDI, have success rates higher than 80%.

• So it is a good idea for a docking program to output multiple docked conformations for analysis.

Table 3. Success Rates of 11 Scoring Functions When Considering Multiple Conformations

	success rate (%) when considering		
scoring function[a]	only the best conformation	the best two conformations	the best three conformations
Cerius2/PLP	76	87	88
SYBYL/F-Score	74	89	90
Cerius2/LigScore	74	78	82
DrugScore	72	82	86
Cerius2/LUDI	67	80	85
X-Score	66	78	79
AutoDock	62	74	78
Cerius2/PMF	52	59	64
SYBYL/G-Score	42	58	66
SYBYL/ChemScore	35	47	51
SYBYL/D-Score	26	45	56

[a] Scoring functions are ranked by their success rates when only the best-scored conformation of each ligand is considered.

Comparison of Docking and Scoring Methods

Docking Accuracy:

• To further evaluate these scoring functions, another test is to classify the 100 complexes into subsets according to the chemical nature of their protein-ligand interactions and then to check the success rate of each scoring function for these subsets.

• For any given protein-ligand complex,

❑ If the contribution of the H-bonding is 50% larger than the hydrophobic contribution, it is classified as the "hydrophilic" type
❑ If the contribution of the hydrophobic term is 50% larger than the H-bonding term, it is classified as the "hydrophobic" type.
❑ Otherwise, the complex is considered to have mixed hydrophilic and hydrophobic factors in the protein-ligand interaction and thus is classified as the "mixed" type.
❑ The X-score empirical function was used to make this classification.

Comparison of Docking and Scoring Methods

Docking Accuracy:

• Generally speaking, higher success rates are observed for the hydrophilic subset.

• Seven scoring functions, i.e., PLP, F-Score, LigScore, DrugScore, LUDI, X-Score, and AutoDock, achieve success rates above 70%.

• This is not surprising because all of these scoring functions have sufficient consideration of hydrogen bonding.

Table 4. Success Rates of 11 Scoring Functions on Different Subsets of Complexes

	success rate (%)			
scoring function[a]	overall (100)	hydrophilic (44)	mixed (32)	hydrophobic (24)
Cerius2/PLP	76	77	78	71
SYBYL/F-Score	74	75	75	71
Cerius2/LigScore	74	77	75	67
DrugScore	72	73	81	58
Cerius2/LUDI	67	75	66	54
X-Score	66	82	59	46
AutoDock	62	73	53	54
Cerius2/PMF	52	68	44	33
SYBYL/G-Score	42	55	34	29
SYBYL/ChemScore	35	32	34	42
SYBYL/D-Score	26	23	28	29

[a] Scoring functions are ranked by their overall success rates.

Comparison of Docking and Scoring Methods

Docking Accuracy:
When the hydrophobic factor in protein-ligand interactions takes a larger share some of these scoring functions perform less satisfactorily such as DrugScore, LUDI, X-Score and AutoDock.

This is also not surprising, since hydrophobic interactions are nonspecific and nondirectional and thus are more difficult to be characterized.

What is surprising is that certain scoring functions, i.e., PLP and F-Score, are able to maintain their success rates across all three subsets.

Table 4. Success Rates of 11 Scoring Functions on Different Subsets of Complexes

	success rate (%)			
scoring function[a]	overall (100)	hydrophilic (44)	mixed (32)	hydrophobic (24)
Cerius2/PLP	76	77	78	71
SYBYL/F-Score	74	75	75	71
Cerius2/LigScore	74	77	75	67
DrugScore	72	73	81	58
Cerius2/LUDI	67	75	66	54
X-Score	66	82	59	46
AutoDock	62	73	53	54
Cerius2/PMF	52	68	44	33
SYBYL/G-Score	42	55	34	29
SYBYL/ChemScore	35	32	34	42
SYBYL/D-Score	26	23	28	29

[a] Scoring functions are ranked by their overall success rates.

Comparison of Docking and Scoring Methods

Docking Accuracy:

- In this docking test, the six relatively successful scoring functions, compared to the AutoDock scoring function, are all empirical scoring functions except DrugScore.

- They typically have well-balanced contributions of polar and nonpolar, enthalpic and entropic factors in protein-ligand binding.

- Another common feature shared by these scoring functions is that they are all calibrated with various sets of protein-ligand complexes.

Comparison of Docking and Scoring Methods

Docking Accuracy:

• The slightly inferior performance of LUDI and X-Score in this test can be understood because, unlike the other four, they are originally developed to reproduce the binding affinities of protein-ligand complexes rather than their structures.

• For example, both LUDI and X-Score use very simple distance and angular functions in their equations, which are based more on chemical intuition rather than a statistical analysis of a large number of experimental structures.

• Moreover, we point out that the hydrophobic term in these two scoring functions needs to be largely improved because the overall performance of these two scoring functions are pulled back by their relatively poor performance in the hydrophobic and the mixed subsets.

Comparison of Docking and Scoring Methods

Docking Accuracy:

• DrugScore, which is a knowledge-based potential of mean force approach, also performs very well (success rate = 72%).

• PMF approaches are different from other scoring methods by deriving potentials through interpreting inverse Boltzmann distributions from a large number of experimental structures.

•

Comparison of Docking and Scoring Methods

Docking Accuracy:

- However, Drug-Score uses an equation combining pairwise potentials and molecular surface based potentials.

- The introduction of molecular surfaces is supposed to capture the hydrophobic effect more effectively, which is a common practice witnessed in empirical scoring functions.

- Thus, the boundary between DrugScore and empirical scoring functions is actually blurred.

$$\Delta W = \gamma \sum_{\text{protein}} \sum_{\text{ligand}} \Delta W_{i,j}(r) + (1-\gamma) \times$$

$$\left[\sum_{\text{ligand}} \Delta W_i(\text{SAS},\text{SAS}_0) + \sum_{\text{protein}} \Delta W_j(\text{SAS},\text{SAS}_0) \right]$$

Comparison of Docking and Scoring Methods

Docking Accuracy:

- In comparison, the PMF approach by Muggue et al. yields a lower success rate (52%) in this test.

- According to this approach, protein-ligand interactions are expressed as a sum of pure distance-dependent pairwise potentials.

- Our opinion is that pairwise potentials may not be as effective as surface-based algorithms for describing the hydrophobic effect in protein-ligand binding.

- The observation that Muggue's PMF approach performs more poorly than DrugScore for the hydrophobic and the mixed subsets seems to support this remark.

Comparison of Docking and Scoring Methods

Docking Accuracy:

• Generally speaking, force field based scoring functions, i.e., AutoDock (success rate = 62%), G-Score (success rate = 42%), and D-Score (success rate = 26%), are less successful in this test.

• One frequently overlooked the fact is that classical force fields are typically not developed for describing intermolecular interactions.

Comparison of Docking and Scoring Methods

Docking Accuracy:

• Therefore, truncating the noncovalent part of a force field and then applying it to protein-ligand binding, such as D-Score, is not expected to give very good results, although it was almost the standard practice in early years.

$$
\begin{aligned}
&V(\bar{\mathbf{r}}) \\
&= \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 \\
&+ \sum_{\text{dihedrals}} K_\chi(1 + \cos(n\chi - \delta)) \\
&+ \sum_{\text{nonbonded-pairs},i,j} \frac{q_i q_j}{4\pi e_0 r_{ij}} - \varepsilon_{ij} \left\{ \left(\frac{R_{\min ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min ij}}{r_{ij}} \right)^{6} \right\}
\end{aligned}
$$

Comparison of Docking and Scoring Methods

Docking Accuracy:

• After some special reparametrization, the performance of force field based scoring functions can definitely be improved, such as what has been seen in the case of AutoDock and G-Score.

• However, the hydrophobic effect still cannot be adequately formularized in a force field equation.

•

• One can see that without exception all of the three force field based scoring functions perform more poorly for the hydrophobic subset and the mixed subset.

Comparison of Docking and Scoring Methods

Docking Accuracy:

- Another practical problem associated with force field based scoring functions is the computation of the electrostatic interaction energy.

- To compute this energy, atom-centered partial charges must be assigned to both the protein and the ligand.

- Theoretical derivation of such a charge distribution in the solvent still remains a problem, especially for a large flexible molecule like a protein.

- For ligands, there is a wide spectrum of schemes ranging from very simple empirical methods to high-level ab initio calculations.

- Potential problem: should not the atomic charges on the protein and the ligand be derived by the same method?

Comparison of Docking and Scoring Methods

Docking Accuracy:

- Another problem is the dielectric constant.

- The binding pocket is more or less shielded from the bulk solvent, and thus, the electrostatic microenvironment inside it is supposed to be different from that of the bulk solvent.

- People have been using two, four, eight, or a distance-dependent dielectric constant to compute the electrostatic interactions between the complex.

Comparison of Docking and Scoring Methods

Consensus Scoring:

• Combining multiple scoring functions, known as consensus scoring, was investigated to identify the correct bound conformation of a given ligand from many computer-generated decoys.

• All the possible double and triple combinations of the six relatively successful scoring functions, i.e., F-Score, LigScore, PLP, DrugScore, LUDI, and X-Score.

Comparison of Docking and Scoring Methods

Consensus Scoring:

• Compared to individual scoring functions, whose success rates range from 66% to 76%, double scoring schemes produce success rates between 76% and 80%, while triple scoring schemes produce success rates between 80% and 84%.

• So it is clear that consensus scoring is also generally more effective than single scoring for molecular docking tasks.

Table 5. Success Rates of Various Consensus Scoring Schemes[a]

consensus scoring scheme	success rate (%)
double scoring	
DrugScore + LigScore	80
DrugScore + F-Score	79
DrugScore + LUDI	79
LigScore + PLP	79
LigScore + F-Score	79
LigScore + X-Score	78
DrugScore + PLP	78
LigScore + LUDI	77
PLP + X-Score	77
PLP + LUDI	77
DrugScore + X-Score	77
PLP + F-Score	76
triple scoring	
LigScore + DrugScore + F-Score	84
LigScore + DrugScore + PLP	84
LigScore + DrugScore + LUDI	83
LigScore + PLP + LUDI	82
DrugScore + PLP + F-Score	82
DrugScore + PLP + X-Score	82
LigScore + DrugScore + X-Score	81
LigScore + PLP + F-Score	80
LigScore + PLP + X-Score	80
DrugScore + PLP + LUDI	80

[a] Since F-Score, LUDI, and X-Score have very similar equations and thus may be less complementary to one other, we do not allow any two of them to appear simultaneously in one consensus scoring scheme.

Comparison of Docking and Scoring Methods

Consensus Scoring:

- Another observation is that which scoring functions are actually included in the consensus scoring scheme seems to be less crucial. All of the double scoring schemes give approximately the same level of success rates and so do all of the triple scoring schemes.

- In conclusion, although consensus scoring does not provide a better understanding of protein ligand interactions, our results demonstrate that it is still a practical strategy for obtaining more reliable results in molecular docking studies.

Table 5. Success Rates of Various Consensus Scoring Schemes[a]

consensus scoring scheme	success rate (%)
double scoring	
DrugScore + LigScore	80
DrugScore + F-Score	79
DrugScore + LUDI	79
LigScore + PLP	79
LigScore + F-Score	79
LigScore + X-Score	78
DrugScore + PLP	78
LigScore + LUDI	77
PLP + X-Score	77
PLP + LUDI	77
DrugScore + X-Score	77
PLP + F-Score	76
triple scoring	
LigScore + DrugScore + F-Score	84
LigScore + DrugScore + PLP	84
LigScore + DrugScore + LUDI	83
LigScore + PLP + LUDI	82
DrugScore + PLP + F-Score	82
DrugScore + PLP + X-Score	82
LigScore + DrugScore + X-Score	81
LigScore + PLP + F-Score	80
LigScore + PLP + X-Score	80
DrugScore + PLP + LUDI	80

[a] Since F-Score, LUDI, and X-Score have very similar equations and thus may be less complementary to one other, we do not allow any two of them to appear simultaneously in one consensus scoring scheme.

Comparison of Docking and Scoring Methods

Binding Affinity Prediction:

- Predicting the correct binding mode of a ligand is only one aspect of molecular docking.

- An equally important aspect of a scoring function is how well it can predict real binding affinities.

- All of the 11 scoring functions were examined to see the correlations between their scores and the experimentally measured binding affinities of the 100 protein-ligand complexes in the test set.

Comparison of Docking and Scoring Methods

Binding Affinity Prediction:

- The performance of these scoring functions in this test is generally less encouraging than their performance in the previous docking test.
- Among all the scoring functions, X-Score gives the best agreement between its scores and the experimental binding affinities with a correlation coefficient of 0.66.
- PLP, DrugScore, and G-Score rank at the second, third, and fourth places, respectively, with correlation coefficients ranging between 0.57 and 0.59.

Table 6. Correlations between Binding Scores and Experimentally Determined Binding Affinities Given by 11 Scoring Functions

	Spearman correlation coefficient (r_s) based on	
scoring function[a]	the experimentally observed conformations	the best-scored conformations
X-Score	0.660	0.698
Cerius2/PLP	0.592	0.607
DrugScore	0.587	0.601
SYBYL/G-Score	0.569	0.531
SYBYL/D-Score	0.475	0.488
SYBYL/ChemScore	0.431	0.435
Cerius2/LUDI	0.430	0.456
Cerius2/PMF	0.369	0.367
Cerius2/LigScore	0.363	0.418
SYBYL/F-Score	0.283	0.253
AutoDock	0.141	0.423

[a] Scoring functions are ranked by correlation coefficients that are calculated by using the experimentally observed conformation of each ligand.

Comparison of Docking and Scoring Methods

Analysis of the Outliers

There are 7 complexes in their test set
for which none of the 11 scoring functions
is able to pick out the correct conformation within an
rmsd threshold of 2.0 Å.

An analysis of these protein-ligand complexes may help to reveal
the shortcomings embedded in today's scoring functions.

Comparison of Docking and Scoring Methods

Among these outliers, are 3 complexes formed between chloramphenicol and type III chloramphenicol acetyltransferases.

In these three complex structures, one remarkable feature is that an entire layer of water molecules exist on the protein-ligand binding interface.

None of the H-bonding groups on the ligand is in direct contact with the protein.

Instead, their interactions with the protein are mediated by some water molecules.

The positions of those water molecules are conserved in all of the three complex structures.

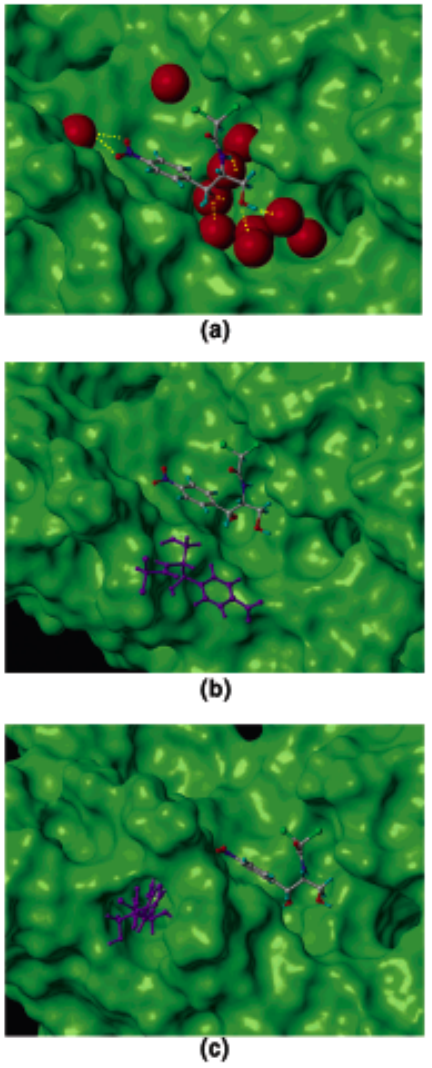


Figure 5. Type III chloramphenicol acetyltransferase in complex with chloramphenicol (PDB entry 3CLA). Chloramphenicol is shown in CPK color with ball-and-stick model. (a) Water molecules on protein–ligand binding interface are shown in red with space-filling model. Dashed yellow lines represent possible H-bonds. (b) Predicted bound conformation by F-Score (in violet, rmsd = 11.1 Å). (c) Predictied bound conformation by DrugScore, LigScore, and PLP (in violet, rmsd = 12.7 Å).

Comparison of Docking and Scoring Methods

In the study all of the water molecules are removed from complex structures because none of the 11 scoring functions can really handle such water-mediated protein-ligand interactions.

After the removal of those water molecules, the experimentally observed conformation is not likely to be favored because it is somewhat suspended in the binding pocket.

Instead, those scoring functions tend to find other locations for the ligand molecule where it can form direct interactions with the protein.

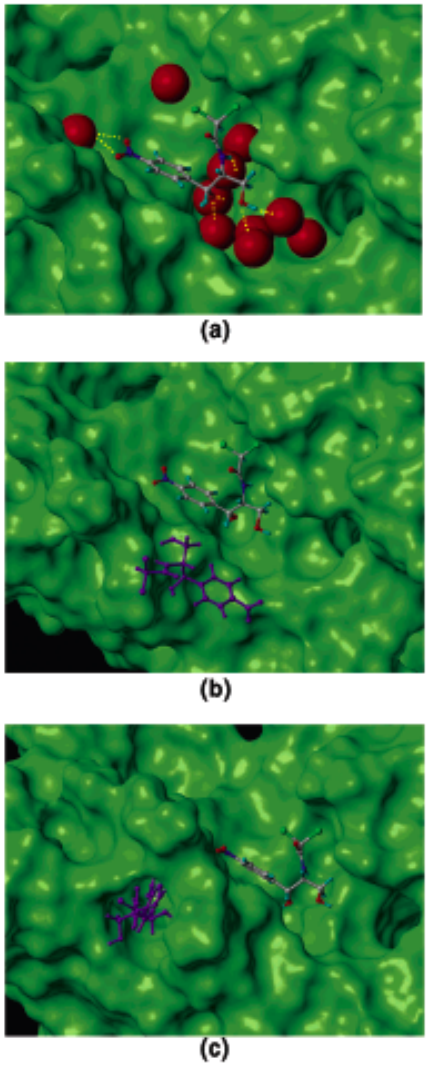


Figure 5. Type III chloramphenicol acetyltransferase in complex with chloramphenicol (PDB entry 3CLA). Chloramphenicol is shown in CPK color with ball-and-stick model. (a) Water molecules on protein–ligand binding interface are shown in red with space-filling model. Dashed yellow lines represent possible H-bonds. (b) Predicted bound conformation by F-Score (in violet, rmsd = 11.1 Å). (c) Predictied bound conformation by DrugScore, LigScore, and PLP (in violet, rmsd = 12.7 Å).

Comparison of Docking and Scoring Methods

For example, the best-scored conformation predicted by F-Score is shown in Figure 5b.

This conformation is not quite native-like because it is not even bound in a cavity.

The best-scored conformation predicted by DrugScore, Lig-Score, and PLP is shown in Figure 5c.

This one is interesting in the sense that the ligand is placed inside a small hole. However, as revealed in the crystal complex structure, that hole is filled with water molecules and is not an alternative binding pocket.

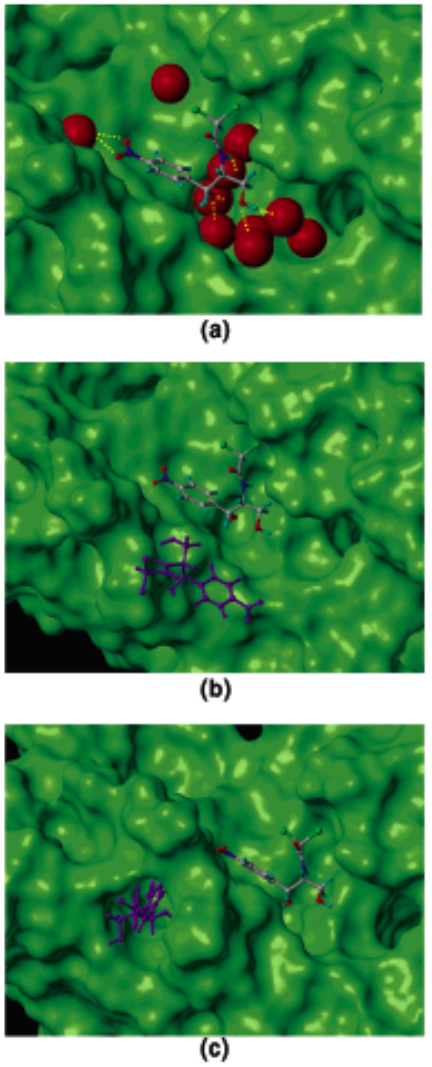


Figure 5. Type III chloramphenicol acetyltransferase in complex with chloramphenicol (PDB entry 3CLA). Chloramphenicol is shown in CPK color with ball-and-stick model. (a) Water molecules on protein–ligand binding interface are shown in red with space-filling model. Dashed yellow lines represent possible H-bonds. (b) Predicted bound conformation by F-Score (in violet, rmsd = 11.1 Å). (c) Predictied bound conformation by DrugScore, LigScore, and PLP (in violet, rmsd = 12.7 Å).

Comparison of Docking and Scoring Methods

Conclusions:

• Among all the scoring functions tested, F-Score, LigScore, PLP, LUDI, DrugScore, and X-Score are able to identify the experimentally observed conformation among a large number of computer-generated decoys for 66-76% of the complexes in the test set.

• Considering the remarkable diversity presented in the test set, this level of success rate is impressive.

• Moreover, combining any two or three of these six scoring functions into a consensus scoring scheme further improves the success rate to nearly 80% or even higher.

Comparison of Docking and Scoring Methods

Conclusions:

- These results suggest that, given an adequate conformational sampling, the performance of today's best scoring functions is totally acceptable for molecular docking tasks.

- Thus, one may want to reexamine the notion that scoring function is the primary problem in molecular docking.

- The tests reveal that binding affinity prediction remains a serious problem.

- For the 100 complexes in the test set, only X-Score, DrugScore, PLP, and G-Score give moderate correlations between their binding scores and experimentally determined protein-ligand binding affinities.

- Unable to predict binding affinities accurately will be a major problem for virtual database screening because true hits may still be missed even when they are correctly docked.