Drug Design

Comparison of Docking and Scoring Methods

- The primary question all docking programs try to address is what combination of orientation and conformation (pose) is the most favourable relative to all the other combinations sampled.

- When applied to screening, the process also requires a comparison of the best pose (or top few poses) of a given ligand with those of the other ligands such that a final ranking can be obtained.

Comparison of Docking and Scoring Methods

Perola et al., Proteins, 2004

Three main topics:

- the importance of test set selection
- the appropriate choice of an accurate docking tool
- the performance of various docking/scoring combinations.

Comparison of Docking and Scoring Methods

The importance of test set selection

• The first part of the study describes the generation of a database of pharmaceutically relevant protein–ligand complexes, which we view as a prerequisite for the evaluation of docking and scoring tools dedicated to drug discovery.

• There is a need for the generation of larger test sets, selected with consistent criteria and highly refined. A careful analysis of the recently reported test sets shows that, while the diversity and pharmaceutical relevance of the protein structures are satisfactory, the same cannot be said with respect to the ligands.

Comparison of Docking and Scoring Methods

The importance of test set selection

- Structural classes that are of less relevance to drug discovery programs (peptides, sugars, nucleotides) are still over-represented, with a high degree of redundancy, and the molecular weight of the ligands generally ranges from 100 to 1000, far beyond the range of interest for a drug discovery program.

- In general the reported test sets only contain a small percentage of truly drug-like ligands. Since such test sets are used in the evaluation/calibration of tools for drug design, it is important that the complexes be representative of what is relevant to the process.

- If the ultimate objective is to predict binding of drug-like molecules to pharmaceutically relevant proteins, complexes between such partners should clearly be emphasized.

Comparison of Docking and Scoring Methods

The importance of test set selection

A new test set of complexes of known binding affinity, geared toward drug-like ligands and suitable for a variety of tasks has been generated:

- evaluation of docking programs and existing scoring functions
- development and calibration of new scoring functions
- analysis of various aspects of protein –ligand binding.

Comparison of Docking and Scoring Methods

The appropriate choice of an accurate docking tool

- The GOLD, ICM and Glide programs were tested for their ability to reproduce crystallographic binding orientations.

- Critical features evaluated include the impact of the nature of the binding site on the accuracy of each program.

Comparison of Docking and Scoring Methods

The performance of various docking/scoring combinations in virtual screening.

- The empirical functions ChemScore, GlideScore and the OPLS-AA force field function are compared.

Comparison of Docking and Scoring Methods

A set of over 200 protein–ligand complexes was initially selected from the Protein Data Bank (PDB) and from the Vertex Pharmaceuticals structure collection according to the following criteria:

General:

° binding constant (Ki or Kd) available

° noncovalent binding between ligand and protein

° crystallographic resolution < 3.0 Å

Ligands:

° molecular weight between 200 and 600

° 1 to 12 rotatable bonds

° drug/lead-like

° structurally diverse

Proteins:

° multiple classes

° diverse within classes

° relevant to drug discovery

Comparison of Docking and Scoring Methods

The initial selection was pruned based on a number of additional criteria.

• In order to prioritize structures that are of higher pharmaceutical relevance, **complexes involving ligand or protein classes** that are less likely to be the **focus of a modern drug discovery program** were excluded.

• Each ligand was included only once, thus avoiding common redundancies. The purpose was to avoid repetitions of almost identical sets of interactions, thus maximizing the diversity of the interactions represented in the test set.

• The final selection included 100 complexes
from the PDB and 50 complexes from the Vertex structure collection.

Comparison of Docking and Scoring Methods

Docking Studies

- The test set of complexes described above was used in the evaluation.

- Each ligand was docked back into the corresponding binding site, and the accuracy of each prediction was assessed on the basis of the root-mean square deviation (RMSD) between the coordinates of the heavy atoms of the ligand in the top docking pose and those in the crystal structure.

Comparison of Docking and Scoring Methods

ICM:

• The Internal Coordinate Mechanics (ICM) program is based on a stochastic algorithm that relies on global optimization of the entire flexible ligand in the receptor field (flexible ligand/grid receptor approach).

• Global optimization (Monte Carlo) is performed in the binding site such that both the intramolecular ligand energy and the ligand– receptor interaction energy are optimized.

• The program combines large-scale random moves (torsional or positional) with gradient local minimization and a history mechanism that both expels from the unwanted minima and promotes the discovery of new minima.

• Five potential maps (electrostatic, hydrogen bond, hydrophobic, van der Waals attractive and repulsive) are calculated for the receptor.

Comparison of Docking and Scoring Methods

Glide :

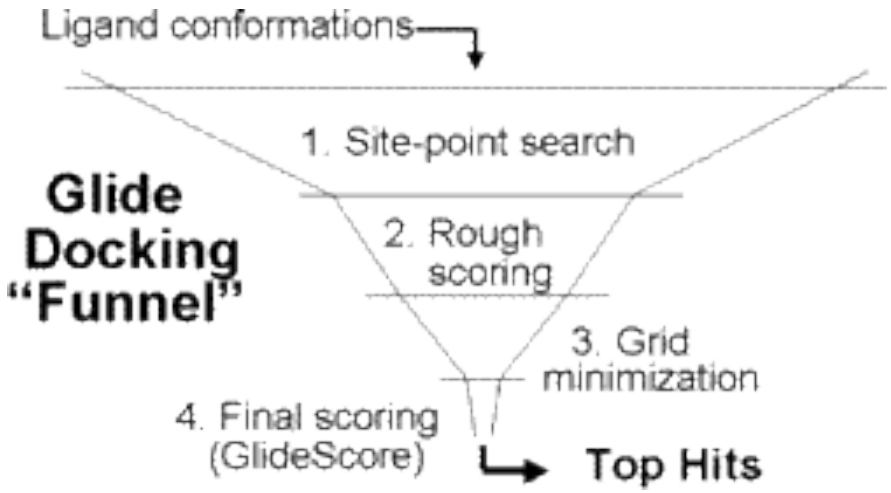
- The Glide (Grid-Based Ligand Docking With Energetics) algorithm approximates a systematic search of positions, orientations, and conformations of the ligand in the receptor binding site using a series of hierarchical filters.

- The shape and properties of the receptor are represented on a grid by several different sets of fields that provide progressively more accurate scoring of the ligand pose. The fields are computed prior to docking.

Comparison of Docking and Scoring Methods

Glide :

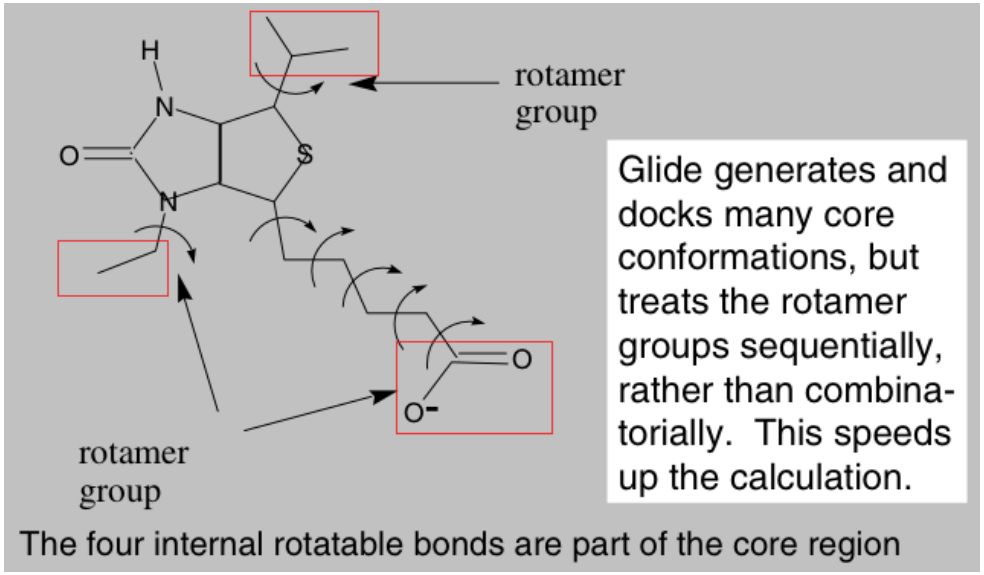
• The search begins with a rough positioning and scoring phase that significantly narrows the search space and reduces the number of poses to be further considered



Comparison of Docking and Scoring Methods

Glide :

- A set of initial ligand conformations is generated through exhaustive search of the torsional minima, and the conformers are clustered in a combinatorial fashion.
- Each cluster, characterised by a common conformation of the core and an exhaustive set of side-chain conformations, is docked as a single object in the first stage.



- It is followed by energy optimisation on an OPLS-AA grid for a few hundred surviving candidate poses.

Comparison of Docking and Scoring Methods

Glide

- The very best candidates are further refined by a Monte Carlo sampling of pose conformation in which torsional minima are examined and the orientation of peripheral groups of the ligand is refined.

- The energy function GlideScore combines empirical and force-field - based terms. It is a more sophisticated version of ChemScore with force field–based components and additional terms accounting for solvation and repulsive interactions

Comparison of Docking and Scoring Methods

GOLD:

- The GOLD (Genetic Optimization for Ligand Docking) program uses a genetic algorithm (GA) to explore the fullrange of ligand conformational flexibility and the rotational flexibility of selected receptor sidechains.

- The mechanism for ligand placement is based on fitting points.

- The program adds fitting points to hydrogen-bonding groups on protein and ligand, and maps acceptor points in the ligand on donor points in the protein and vice versa.

- Additionally, GOLD generates hydrophobic fitting points in the protein cavity onto which ligand CH groups are mapped.

Comparison of Docking and Scoring Methods

• The genetic algorithm optimizes flexible ligand torsions, ligand ring geometries, torsions of protein OH and NH3 groups, and the mappings of the fitting points.

• The docking poses are ranked based on a molecular mechanics–like scoring function, which includes a hydrogen- bond term, a 4-8 intermolecular van der Waals term, and a 6-12 intramolecular van der Waals term for the internal energy of the ligand.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: General performance

The results of this study clearly identified Glide as the most accurate of the three docking programs examined, with 61% of the top-ranking poses within 2.0 Å of the corresponding crystal structure.

Both GOLD and ICM also performed reasonably well, with 48% and 45% of top ranking poses meeting the same criterion, respectively.

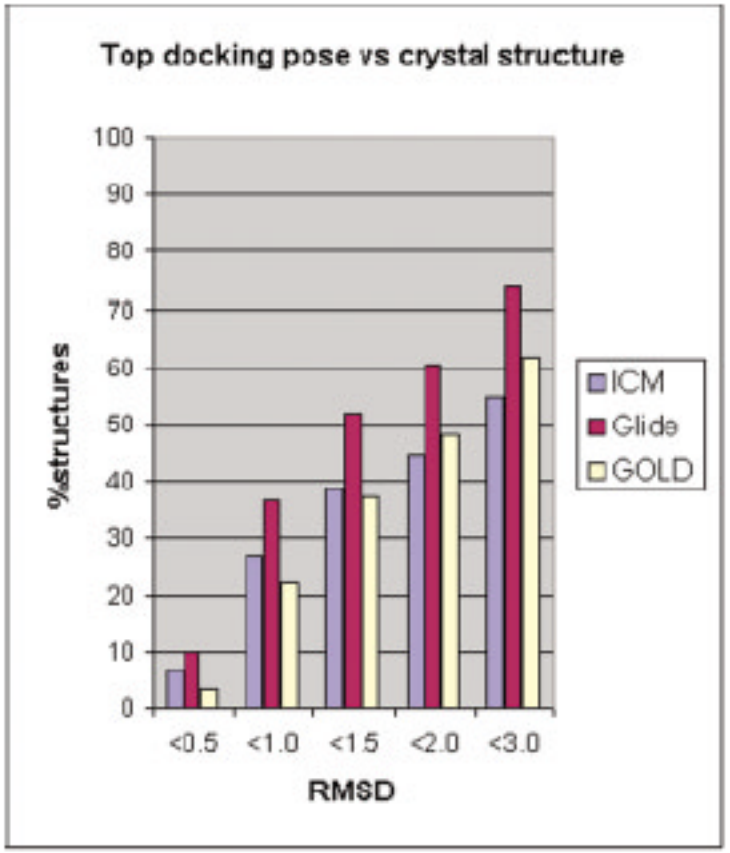


Fig. 1. Distribution of the RMSDs between the top-ranked docking poses and the corresponding crystal structures. The RMSDs were calculated on the coordinates of the heavy atoms of the ligands. *x* axis: RMSD cutoffs; *y* axis: percentage of top-ranked docking poses within a given RMSD cutoff from the crystallographic pose.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: General performance

Analysis of the top 20 docking solutions shows that GOLD was generally as effective as Glide in sampling the correct pose and placing it in the top 20.

When the top 20 docking poses were compared to the corresponding crystal structure, the percentages of best poses (lowest RMSD) within 2.0 Å from the experimental structure were 79% for Glide and 77% for GOLD.

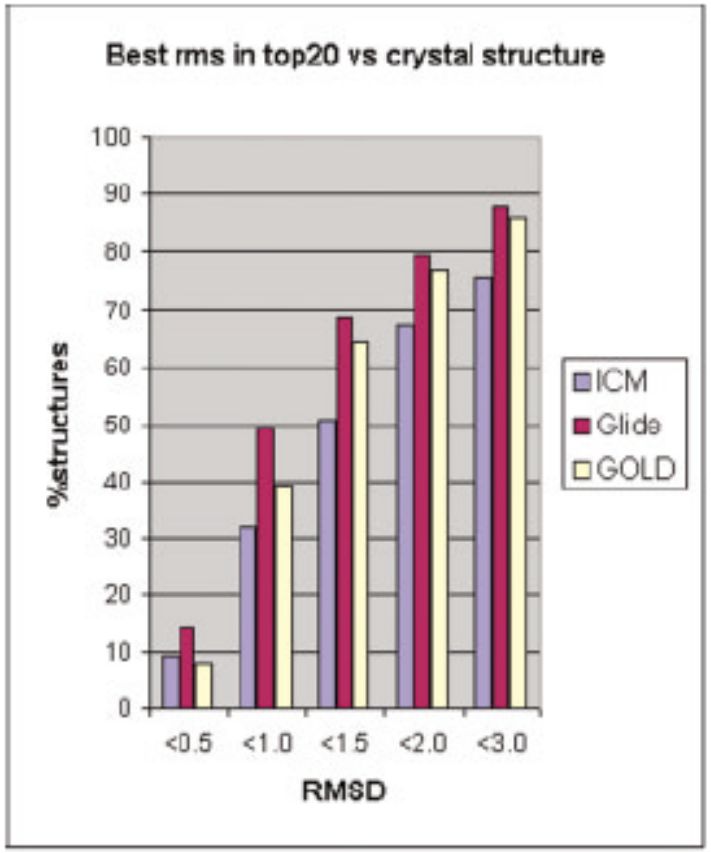


Fig. 2. Distribution of the RMSDs between the closest of top 20 docking poses (lowest deviation) and the corresponding crystal structure for each complex. *x* axis: RMSD cutoffs; *y* axis: percentage of closest docking poses within a given RMSD cutoff from the crystallographic pose.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: General performance

- Based on this observation, the GOLD algorithm appears to be equally efficient in terms of sampling, but the Glide scoring function seems more accurate than the GOLD fitness function in ranking the sampled poses.

- The ICM algorithm appears to perform less well than the other two in terms of sampling.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs:
Correlations between active site features and docking accuracy

• In order to assess where the difference between Glide and the other programs lies and on what kinds of systems each program performs best, the dependence of the docking accuracy on specific structural descriptors was analyzed.

• The complexes were classified in a binary or ternary fashion with respect to three structural features: flexibility of the ligand, predominant nature of the interactions between ligand and receptor, and degree of solvent exposure of the binding pocket.

• Statistical analysis of the docking accuracies was performed with regard to such features.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs:
Correlations between active site features and docking accuracy

- In terms of flexibility, it is well known that the accuracy of any docking program decreases with the number of rotatable bonds of the ligand.

- The size of the conformational space to be sampled increases exponentially with ligand flexibility, and the thoroughness of the sampling has to be partially sacrificed to keep the computing time within reasonable limits.

- Different algorithms use different methods to circumvent the problem and maximize the efficiency of the conformational sampling.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

- The test systems were divided in two groups: 87 complexes of ligands with 1–6 rotatable bonds and 63 complexes of ligands with 7–12 rotatable bonds.

- The results show that the loss of accuracy going from less flexible to more flexible ligands is relatively small for Glide (from 67% to 52% of correct solutions) and much more dramatic for GOLD and ICM, with the latter losing more than half of its predictive power.

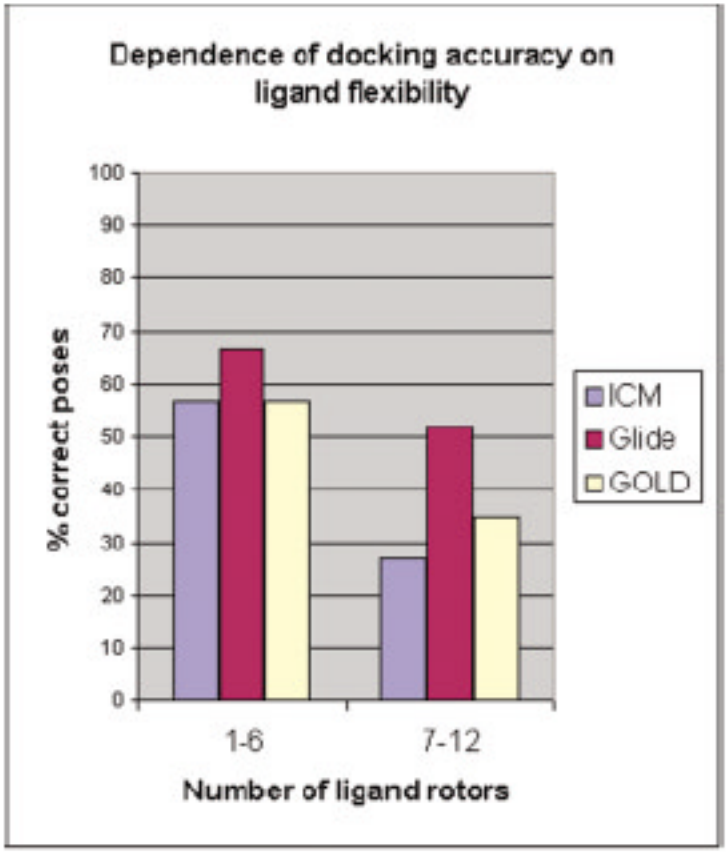


Fig. 4. Performance of the three docking programs on complexes with lower and higher ligand flexibility. *x* axis: range of ligand flexibility; *y* axis: percentage of top-ranked docking poses within 2.0 Å from the corresponding crystal structures.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

• This indicates that the multistage systematic algorithm implemented in Glide results in a more extensive coverage of conformational space than both the genetic algorithm and the stochastic search implemented in GOLD and ICM.

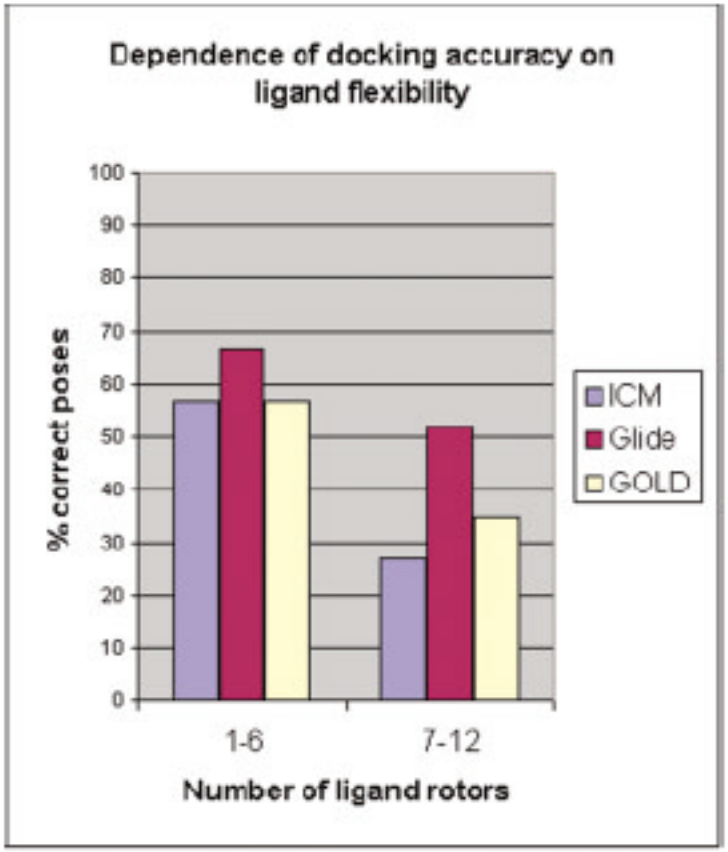


Fig. 4. Performance of the three docking programs on complexes with lower and higher ligand flexibility. *x* axis: range of ligand flexibility; *y* axis: percentage of top-ranked docking poses within 2.0 Å from the corresponding crystal structures.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

• In terms of interactions, hydrogen bonds and hydrophobic interactions are considered the main contributors to protein–ligand binding in the vast majority of complexes.

• In order to divide the complexes in our test set between hydrogen bond–driven and hydrophobic-driven, the number of hydrogen bonds between protein and ligand in each complex was determined.

• The degree of hydrogen bonding (DHB), defined here as the ratio between number of hydrogen bonds and number of heavy atoms in the ligand, was used to define the dominant contributor to binding for each complex.

• Complexes with a DHB of 0.15 or higher were classified as hydrogen bond–driven, while complexes with a DHB of 0.10 or lower were classified as hydrophobic-driven, with the remaining complexes in the intermediate category.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

• The results show that all programs perform best on complexes in which there is a relatively even balance between hydrogen bonding and hydrophobic interactions.

• Interestingly, for both ICM and GOLD, the docking accuracy decreases dramatically when binding is mainly driven by hydrophobic interactions, while Glide, which appears to be somewhat less sensitive to the nature of binding, performs better on hydrophobic-driven complexes than on hydrogen bond–driven complexes.

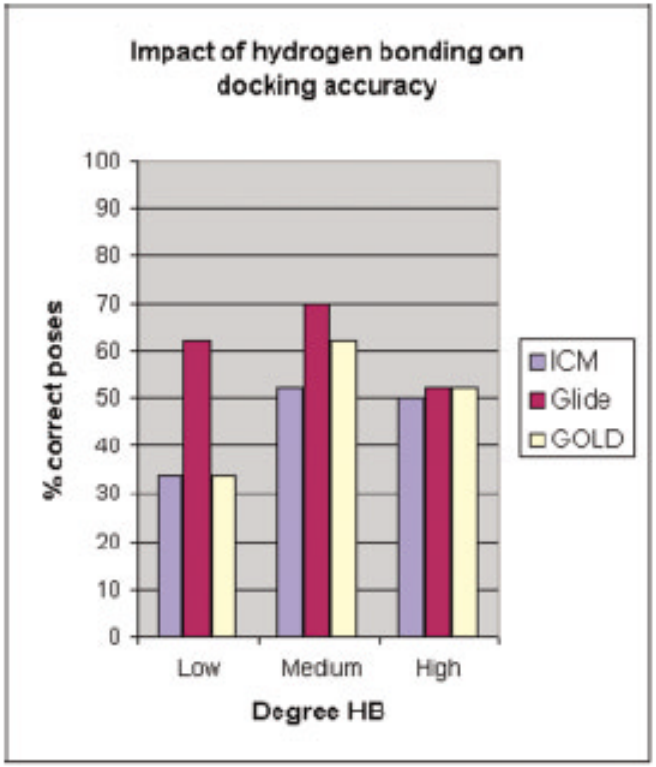


Fig. 5. Performance of the three docking programs on complexes with different degrees of hydrogen bonding between ligand and protein. The degree of hydrogen bonding (DHB) is defined as the ratio between the number of hydrogen bonds between ligand and protein and the number of heavy atoms in the ligand. *x* axis: DHB (low: DHB $\leq$ 0.10; medium: 0.10 $<$ DHB $<$ 0.15; High: DHB $\geq$ 0.15); *y*-axis: percentage of top-ranked docking poses within 2.0 Å from the corresponding crystal structures.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

- The preference of GOLD for complexes rich in hydrogen bonds has been pointed out previously and it can be ascribed to the nature of the algorithm, in which the mapping of hydrogen bond fitting points plays a major role.

- In the case of ICM, this tendency has not been reported; one possible explanation is that in a Monte Carlo search, mostly characterized by low-energy moves, the presence of a set of hydrogen bonds may lock part of the molecule into its correct orientation during the search, thus allowing for a more efficient sampling of the rest of the molecule.

- For Glide, the difference in performance is less significant, and this consistency across active sites with various degrees of hydrophobicity/hydrophilicity is another reason for its better performance on the complete test set.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

• When interactions with metals were specifically considered, no difference in performance was observed among the three programs: on the 24 metal-containing complexes, Glide selected a solution within 2.0 Å of the experimental structure 9 times, while ICM and GOLD succeeded 8 times in the same subset.

• The success rate of the three programs on such systems was significantly poorer if compared to the overall performance, which points to the necessity of further progress in this area, especially considering the continued interest in zinc metalloproteins as drug discovery targets.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

• The third aspect analyzed in this context is the impact of the degree of burial of the binding pocket on the docking accuracy achieved with different search algorithms.

• It is generally the case that buried binding sites restrict the number of orientations, positions, and conformations accessible to putative binders, but at the same time, they require a finer sampling in order to achieve the proper set of interactions without clashes.

• On the other hand, solvent-exposed sites require more extensive sampling to cover all the accessible poses, but at the same time are more tolerant with respect to the combination of pose descriptors required to achieve the proper set of interactions.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

- In this study, the binding sites of the test complexes were divided into three groups, with low, medium, or high degree of burial, and the docking results were dissected accordingly.

- In order to assign the complexes to each group, the solvent-accessible surface area of the crystallographic ligand was calculated in the presence and in the absence of the bound protein partner, and the fraction of buried ligand was determined for each complex.

- The degree of burial was defined as low if the fraction was 0.75 or lower, high if the fraction was 0.90 or higher, and medium for values in between.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

The analysis of the performances attained by the three programs on each class, shows that all of them achieve the highest degree of accuracy on complexes with buried binding pockets, and consistently lose accuracy with an increase in solvent exposure.

Once again, Glide appears to be relatively less sensitive to the features of the binding pockets, while ICM shows the largest decay in performance going from buried to solvent-exposed pockets.

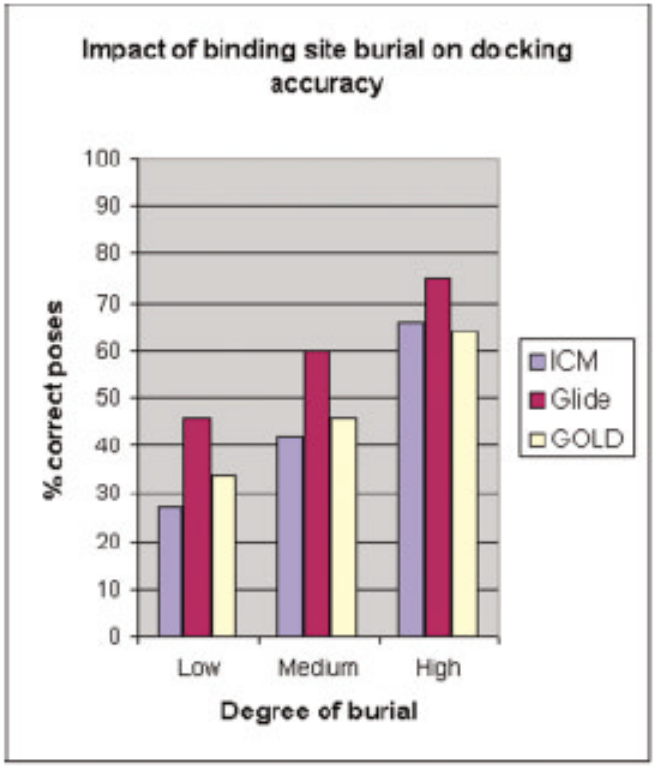


Fig. 6. Performance of the three docking programs on complexes with different degrees of binding site burial. The degree of burial is defined as the fraction of the solvent-accessible surface area of the ligand that becomes buried upon binding. *x* axis: degree of burial; *y* axis: percentage of top-ranked docking poses within 2.0 Å from the corresponding crystal structures.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Correlations between active site features and docking accuracy

- These results indicate that all three search algorithms can explore an enclosed binding site much more efficiently than a relatively open one, and also points to the obvious observation that, in a more sterically constrained site, the best pose for a given ligand is more unequivocally defined by the shape of the site.

- As a consequence, the likelihood of generating multiple poses with similar score is much lower and the selection of the best pose is more straightforward.

- For the same reason, it is safe to say that, when docking compounds in a buried binding pocket, an efficient sampling process may be more important than an accurate scoring/ranking method, while in a solvent-exposed pocket, both aspects become equally important.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Analysis of problematic structures

- In addition to the general trends observed, the results of this study highlight some limitations and shortcomings that are common to all docking programs examined.

- In 12 cases, none of the top 20 poses generated by any of the three programs was within 2.0 Å of the experimental structure.

- Most of these common failures can be ascribed to a combination of structural features.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Analysis of problematic structures

- Four of the problematic complexes are characterised by a dominance of hydrophobic interactions in solvent-exposed sites.

- In such cases, the shape of the pocket does not help to restrict the number of possible binding orientations, and the lack of a set of specific anchoring points for the ligand makes the selection of the best pose very challenging.

- Moreover, all four ligands are relatively flexible (8–9 rotatable bonds) which adds to the sampling problem.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Analysis of problematic structures

- Three complexes present highly flexible ligands in almost completely buried binding sites.

- In these cases the tightness of the binding pockets and the very specific conformational requirements for the ligands to achieve the correct pose call for a very thorough sampling process, which is very hard to attain within the boundaries of a limited computing time.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Analysis of problematic structures

- Another aspect that is sometimes problematic in docking is the presence of charged functionalities in the ligand, because the desolvation energy required for such groups to become available for interaction with the protein is overlooked by most scoring functions.

- In two of the failed complexes there is a basic amine in the ligand that does not interact with any protein residue when the crystal complex is analyzed. Docking functions tend to favour poses in which such groups form hydrogen bonds and/or salt bridges.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Analysis of problematic structures

• Docking accuracy can also be impaired by the occurrence of unconventional interactions, not properly parameterized in the fitness functions of the programs employed.

• Two examples are hydrogen bonds between hydrogens of electron-poor aromatic rings and protein acceptors
and hydrogen bonds between the imino form of anilino nitrogens and protein donors.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Summary of relevant findings

- The Glide program is shown to have the highest degree of accuracy on a wide and diverse set of systems, which makes it the tool of choice in most cases.

- The Glide program is more tolerant than both ICM and GOLD of the increase of ligand flexibility, which seems to point to a more effective conformational sampling method.

- Analogously, Glide appears to be less sensitive to variations in the polarity of the binding pocket, with a slight preference for complexes with prevalent hydrophobic character but a solid performance across the board.

- On the other hand, ICM and GOLD can be considered as reliable as Glide when operating on highly polar binding sites, where binding is strongly driven by hydrogen bonding.

Comparison of Docking and Scoring Methods

Evaluation of Docking Programs: Summary of relevant findings

- Comparatively, the ability of ICM and GOLD to predict complexes where binding is driven by hydrophobic interactions is relatively poor.

- All three programs perform best on buried binding pockets, with a gradual decrease in performance at the increase of solvent exposure.

- In general, some systems remain a challenge for docking at the current stage, which suggests that there is still a margin for improvement on the existing methods.

- In particular, the inclusion of properly weighted solvation terms and a more effective representation of metal-mediated interactions in the fitness functions appear to be highly desirable.