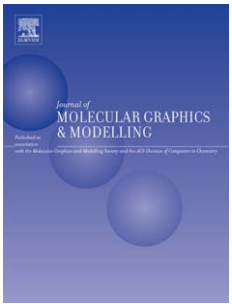


Title: eHiTS: A new fast, exhaustive flexible ligand docking system

Authors: Zsolt Zsoldos, Darryl Reid, Aniko Simon, Sayyed Bashir Sadjad, A. Peter Johnson

PII: S1093-3263(06)00099-4
DOI: doi:10.1016/j.jmgm.2006.06.002
Reference: JMG 5589

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 1-2-2006
Revised date: 8-6-2006
Accepted date: 12-6-2006

Cite this article as: Z. Zsoldos, D. Reid, A. Simon, S.B. Sadjad, A.P. Johnson, eHiTS: A new fast, exhaustive flexible ligand docking system, *Journal of Molecular Graphics and Modelling* (2006), doi:10.1016/j.jmgm.2006.06.002

eHiTS: a new fast, exhaustive flexible ligand docking system

Zsolt Zsoldos*, Darryl Reid, Aniko Simon,
Sayyed Bashir Sadjad, A. Peter Johnson†
SimBioSys, Inc.‡, 135 Queen's Plate Drive, Suite 520,
Toronto, On., M9W 6V1, Canada

June 8, 2006

Abstract

The flexible ligand docking problem is divided into two subproblems: pose/conformation search and scoring function. For successful virtual screening the search algorithm must be fast and able to find the optimal binding pose and conformation of the ligand. Statistical analysis of experimental data of bound ligand conformations is presented with conclusions about the sampling requirements for docking algorithms.

eHiTS is an exhaustive flexible docking method that systematically covers the part of the conformational and positional search space that avoids severe steric clashes, producing highly accurate

*To whom all correspondence should be addressed, zsolt@simbiosys.ca
†Chemistry Department, The University of Leeds, Leeds, LS2 9JT, U.K.
‡http://www.simbiosys.ca/

docking poses at a speed practical for virtual high throughput screening.

The customizable scoring function of eHiTS combines novel terms (based on local surface point contact evaluation) with traditional empirical and statistical approaches.

Validation results of eHiTS are presented and compared to three other docking software on a set of 91 PDB structures that are common to the validation sets published for the other programs.

1 Introduction

Structure based drug design is well established as a key component in the drug discovery process for many pharmaceutical companies. The screening of large libraries of compounds against targets in search of novel scaffolds suitable for further lead refinement, is relatively commonplace. However, the high cost of experimental screening prompts the use of computational (virtual) screening techniques as a preliminary filter to reduce the size of the library prior to the much more expensive experimental screening phase. One such computational screening technique is flexible ligand docking, where the candidate ligands are fitted to the 3D structure of the target receptor with allowance for the conformational flexibility of the ligands.

The process of docking a compound into a receptor site can be computationally demanding. It can be viewed as an energy minimization problem, however most of the available molecular mechanics programs are too sensitive to local minima to find the appropriate docking poses[14]. Various stochastic search methods exist which attempt to solve this problem, including Simulated Annealing (e.g. AutoDock2[9], Dockvision[10],MCDOCK[15]), Genetic Algorithms (GOLD[12], AutoDock3[17], DockVision[10]), Tabu Search (ProLeads[24]), etc.

They have been reported to be successful in reproducing the experimental binding conformations of some ligand receptor complexes[12]. The search algorithm in these methods is a random probing technique, driven solely by a scoring function. Consequently, these methods do not guarantee a systematic coverage of the search space.

Systematic alternatives to these random trial and error approaches do exist, including incremental construction based (FlexX[22], Hammerhead[23], DOCK4) and multiple conformer rigid body docking (e.g. FLOG[13], DOCK3[6] or FRED[16]). Even though these methods are systematic, unfortunately, they still do not provide an exhaustive search of the conformational and pose space. The incremental construction methods employ a coarse sampling of conformations using a small number of discrete rotomers. The multiple conformer rigid docking systems use a few hundred low energy conformers of the ligand. None of the current docking programs can guarantee screening with no false negatives, which implies that they can miss potential solutions.

As shown later, statistical analysis of experimental data from bound ligand conformations illustrates that sampling of low energy conformers is insufficient to reproduce protein-ligand binding geometries, a much more exhaustive search is required. eHiTS (electronic High Throughput Screening) offers the first truly exhaustive systematic search algorithm that considers all poses without severe steric clash. Employing unique graph matching algorithms and using dock tables, stored in SQL databases, eHiTS is suitable for high-throughput screening applications.

In evaluating the eHiTS algorithm, its accuracy in reproducing known bound conformations will be considered along with the ability to enrich database selections with actives. Accuracy will be measured as the ability of the docking algorithm to replicate the docking poses of ligands from co-crystallized proteins.

2 Pose and conformational sampling requirements

A fundamental design goal of the eHiTS system is to provide an exhaustive systematic search of the part of the conformational and pose space that avoids severe steric clashes with sufficiently fine sampling to reproduce experimentally observed binding modes. In theory, a truly exhaustive search should explore the infinite continuum of rotational and translational space. In practice, discrete sampling is acceptable *if* it is fine enough to *not* miss a solution.

Statistics on hydrogen bond geometry in small molecular crystal structures [11] show a range of 1.6Å to 2.2Å distance between the hydrogen and the acceptor atoms, i.e. it can be described as 1.9Å ± 0.3Å.

Hydrophobic contacts are observed [4] in the range of 3.2Å to 4.2Å between the atom centers of two carbons, i.e. 3.7Å ± 0.5Å.

Aromatic π stacking interactions and metal ion interactions also have their own ranges of acceptable geometry with similar tolerances. It is clear that a half Angstrom difference in atom positions may mean losing a crucial hydrogen bond or cause a severe steric clash instead of a perfect van der Waals contact. Therefore, we define sufficient sampling to mean that atom positions must be sampled at least every half an Angstrom.

This definition of sufficient sampling for atom displacements implies a requirement for rigid fragment rotation and dihedral angle sampling. A simple trigonometric calculation shows that a tangential movement of 0.5Å is caused by rotation of about 5° at a radius of 7Å. Drug-like ligands can easily reach or exceed the size of 7Å, therefore rotations and dihedral angles must be sampled at least every 5 degrees.

The structure shown in Figure 2 is that of the ligand from PDB code 1CX2, and can be used to demonstrate the rotational sampling requirement. The left side of the ligand is anchored by hydrogen bonds. The dihedral angle about

the bond indicated by the red arrow will greatly influence the position of the Fluorine at the top of the figure, because that atom is 7Å away from the axis of rotation. If the dihedral angle changed by 5°, then the Fluorine would move by 0.61Å Angstrom. If a sampling algorithm missed the correct dihedral by 10 or 15 degrees then the Fluorine would end up in the center of a receptor carbon atom causing the most severe steric clash – two atom centers at zero distance – instead of creating a perfect hydrophobic surface contact.In other words, a 15-30+ degree sampling is far too crude to be useful for a docking program that aims to be exhaustive.

Figure 3 shows another practical example, from experimental X-ray crystal structure data (PDB code 1JQY), which demonstrates the inadequacy of using only staggered or gauche conformers in a docking study. The observed zero degree dihedral (indicated by the arrow) between two sp3 carbon atoms would represent a high energy local conformation in a ligand in isolation and there is nothing in the ligand structure which requires or compensates for such a strain.

Bound conformations of 5000 ligands in high resolution (less than 2.5Å) crystal structures from the RCSB Protein Data Bank (PDB) have been analyzed to collect statistical data on the dihedral angles of rotatable bonds. Table 1 shows how many ligands have *all* their dihedral angles within the given ± range to either a staggered or a gauche value, i.e. how many bound ligand conformations would be found within a given error if only those dihedral angles were sampled. Another important data point is that about 10% of the bound ligand conformers exhibit at least one eclipsed dihedral angle, i.e. 0±5° between sp3 centers each bearing one additional heavy-atom neighbor. 97% of X-ray conformations in this set deviate by more than 5° from conformations generated by sampling the dihedrals of each rotatable bond every 60°. It is clear from the data that it is necessary to include conformations which in an isolated molecule would be of high energy, in order to sample the conformations adequately for

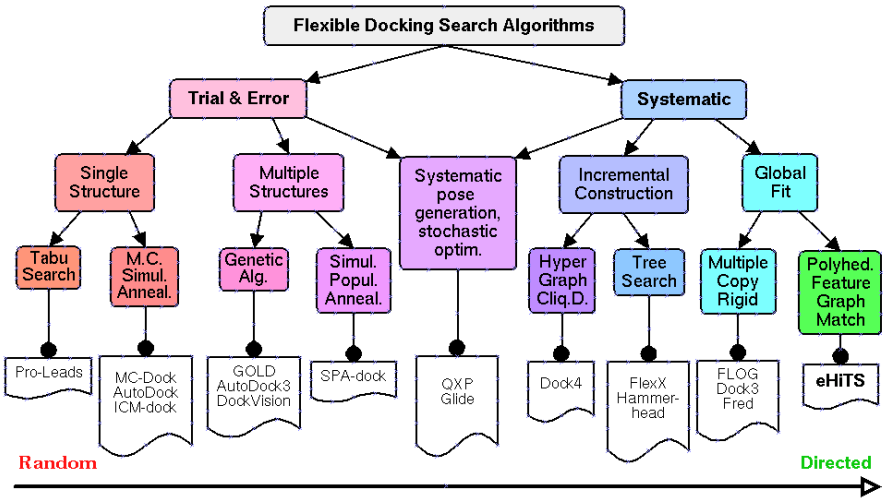


Figure 1: Categorization of flexible docking algorithms

Error limit	Number of ligands	Percentage
±5°	108	2.2%
±10°	211	4.2%
±15°	315	6.3%

Table 1: Statistics concerning staggered and gauche conformers in X-ray structures of bound ligands from 5000 PDB entries with 2.5Å or better resolution. The Table shows how many ligands have *all* their dihedral angles within the given ± range to either a staggered or a gauche value.

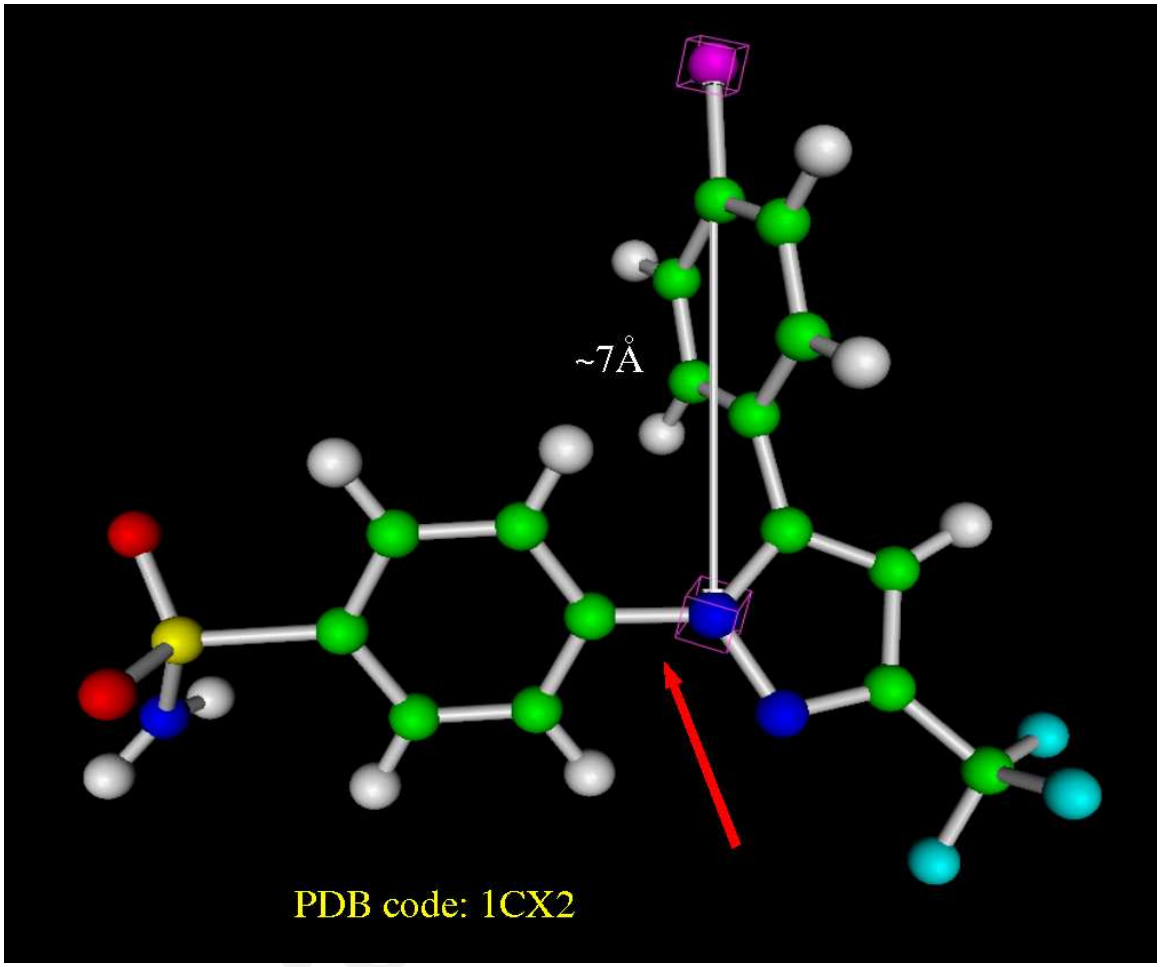


Figure 2: The ligand structure in PDB entry 1CX2. The position of the Fluorine atom at the top of the figure is greatly influenced by the dihedral angle of the bond indicated by the red arrow.

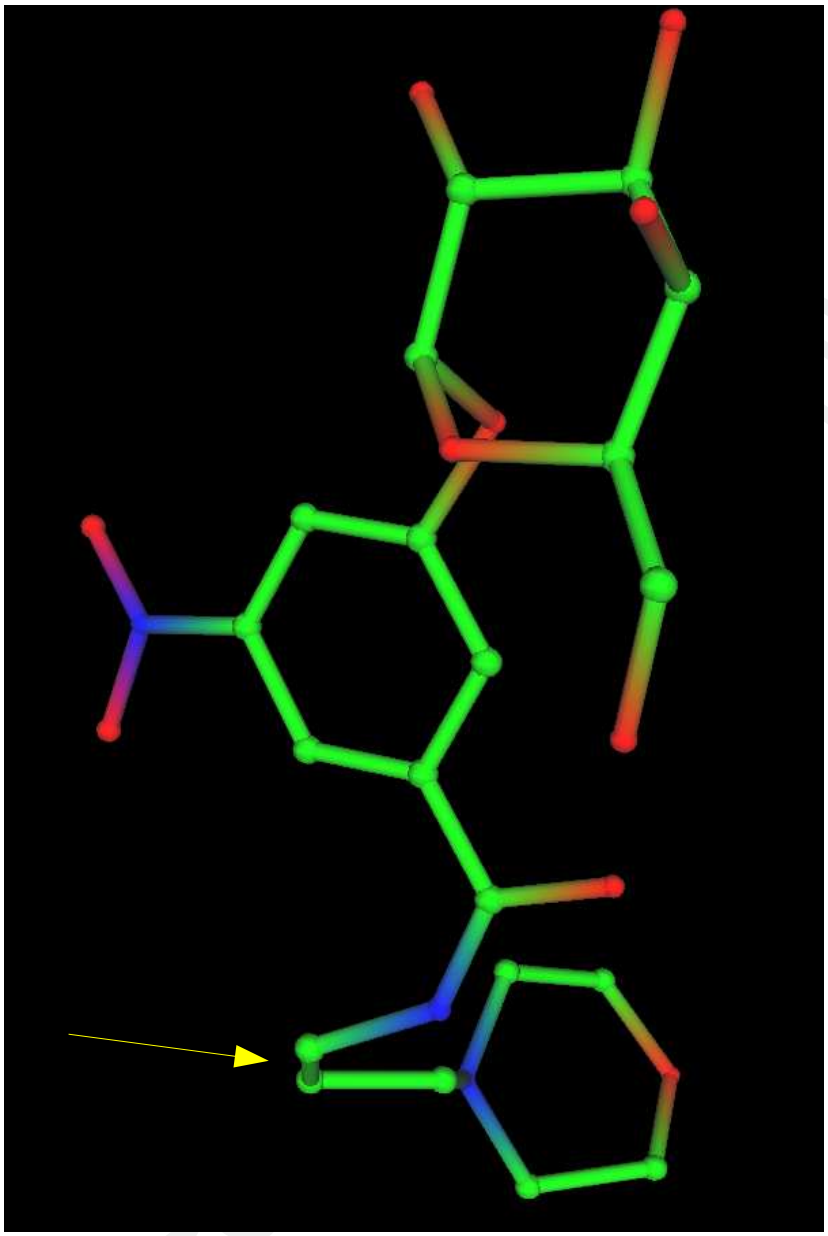


Figure 3: The bound ligand conformation in PDB entry 1JQY.

docking.

Many researchers have performed similar analysis[18, 2, 1, 19] and come to essentially the same conclusions regarding high-energy conformations of bound ligands.

2.1 Search space size

The size of the search space can easily be calculated from the sampling requirement defined above. For an average sized ligand with six rotatable bonds, the following formula is computed:

Translations along 3 axes	$\forall 0.5$ Å in 10Å box:	$(10/0.5)^3 = 20^3$
Orientations about 3 axes	$\forall 5°$ in 360°:	$(360/5)^3 = 72^3$
Dihedral angle sampling	$\forall 5°$ in 360°:	$(360/5)^6 = 72^6$
Total number of poses:	$20^3 \times 72^3 \times 72^6 \approx$	2×10^{20}

This number (ten to the power twenty) is so huge that brute force evaluation of all those poses with a relatively fast scoring function – that can process 2 thousand poses per second – would take 3 billion years on a single CPU. Using the largest current supercomputer (BlueGene/L in California DOE/NNSA/LLNL with 131 thousand CPUs, according to www.top500.org at the time of writing) it would still take more than 20 thousand years to dock a single ligand.

Stochastic methods that employ fine enough sampling, do search this same vast space, but instead of systematic sampling, they employ random walks. Decisions are made based on a goal function evaluation and some stochastic decision process whether or not to keep a given trial pose. However, new trial poses are selected by some random alteration of an already tested pose. There is *no* driving force employed towards new areas of the search space that are yet unexplored. Therefore the poses examined by stochastic methods, if represented as points in N-dimensional space, are comparable to Brownian movement. Such random walks are known[7] to over-sample some regions while leaving some

large areas completely unexplored. The flexible ligand docking pose space is 10-20 dimensional (depending on the number of rotatable bonds) and the sampling problems of random walks are much more severe than they are in low dimensional problem space.

Our goal was to develop an intelligent exhaustive method that can limit the fine sampling of the search space to areas of interest where good scoring solutions may reside, while eliminating large portions of the vast search space where it is guaranteed that no good scoring position can be found.

3 The eHiTS method

As demonstrated above, brute force evaluation of *all possible poses and conformations* with sufficiently fine sampling is not feasible within practical CPU time limits. Therefore, the search space must be reduced. One reduction applied by eHiTS is to limit the search to conformations and poses that avoid severe steric clashes between receptor and ligand, i.e. where geometric fit is possible.

In order to explore the vast search space exhaustively in an efficient manner, our approach involves sub-division of the task into smaller partial problems that are easier to solve. However, unlike DOCK or FLexX, eHiTS does *not* use an incremental construction method, but instead attempts to find the global optimum by enumerating combinations of independent partial structure dockings.

eHiTS has a novel flexible ligand docking method that is exhaustive on the conformations and poses that avoid severe steric clashes between receptor and ligand. The algorithm generates all major docking modes that are compatible with the steric and chemistry constraints.

First the binding pocket is determined by building a steric grid for the whole receptor, dividing regions into separate pockets and identifying the possible interaction sites. Then, a cavity description is built that consists of thousands

of geometric shapes (polyhedra).

The ligand is divided into rigid fragments and connecting flexible chains. eHiTS docks *all* rigid fragments to *all* possible places in the cavity *independently* of each other. This is *not* an incremental construction, all rigid fragments are docked to every possible place regardless of the other fragments. Although, the poses are scored, no local (biased) decision is made to reject any sterically feasible pose for any rigid fragment based on interaction score.

An exhaustive matching of compatible rigid fragment pose sets is performed by a rapid hyper-graph clique detection algorithm. This may yield a few hundred (small pocket, few rigid fragments) to several million (large pocket, many small rigid fragments) acceptable combinations of poses. However, at this point, the scores for each component have been evaluated, so it is possible to make a *global* decision as to which fragment pose combination is the best.

The flexible chains are then fitted to the specific rigid fragment poses that comprise a matching pose set. The reconstructed solutions define a rough binding pose and conformation of the ligand. These poses are refined by a local energy minimization in the active site of the receptor, driven by the scoring function. Figure 4 shows two snapshots of the docking (overlayed), three fragments in their rigid docked poses in thin-bond representation and the final full ligand pose after optimization in thick-bond drawing.

3.1 Geometric shape and chemical feature graph

The fragmentation of the ligand is focused on separating rigid fragments from the flexible linkers. All ring systems are considered rigid and their conformation is preserved as given in the input. Therefore it is desirable to use multiple ring conformers (e.g. chair, boat and twist boat for a cyclohexane) for complete conformational sampling.

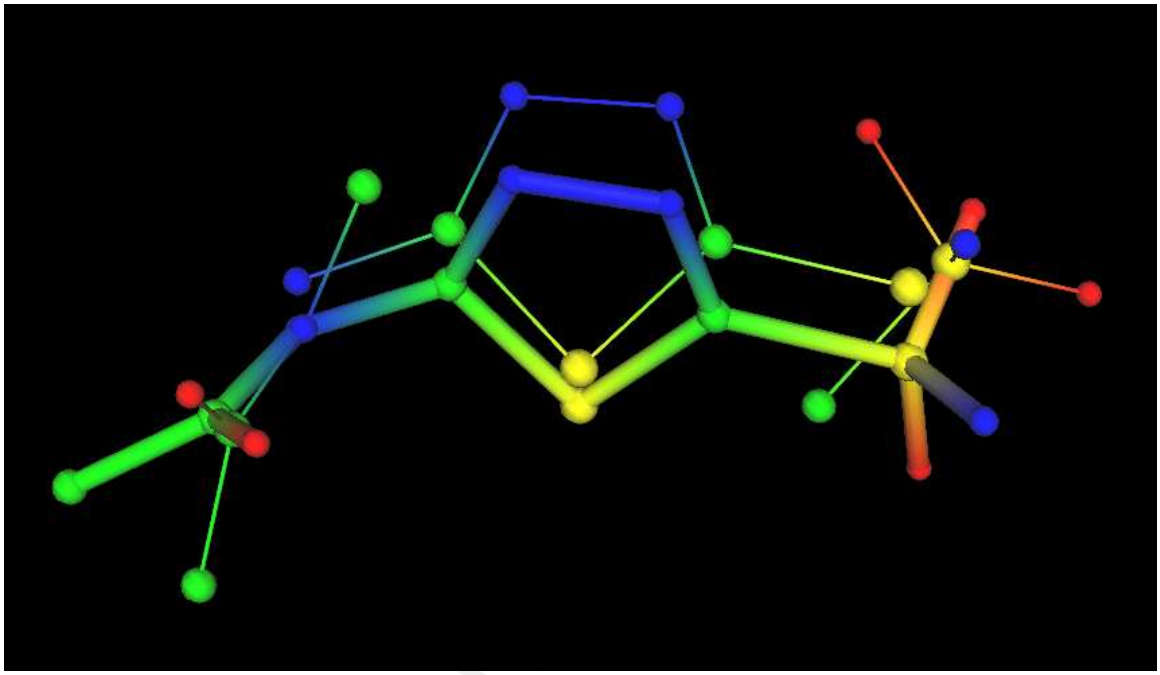


Figure 4: The ligand of PDB code 1azm in two stages of the eHiTS docking. The 3 individual fragment poses (after the rough rigid docking) are shown in thin-bond representation. The thick-bond molecule shows the ligand pose after reconstruction and optimization. The colors represent atom types: green is carbon, blue is nitrogen, red is oxygen, yellow is sulphur.

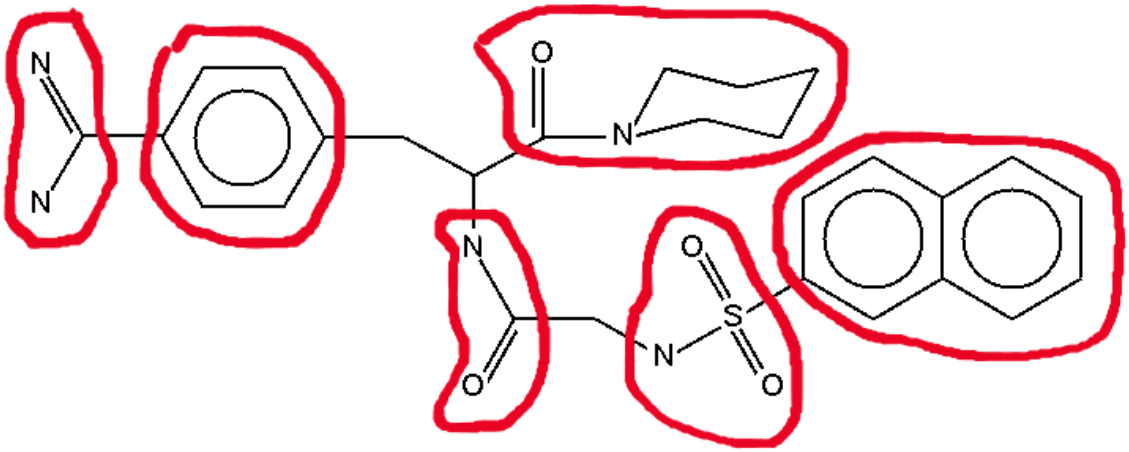


Figure 5: The ligand is broken into rigid fragments and flexible chains

Acyclic fragments with double or normalized (resonance) bonds and sp2 hybridized atoms are also considered rigid, e.g. including the amide functional group. Figure 5 shows an example of the fragmentation of a ligand. Whenever a bond is broken during this fragmentation, both atoms of the bond are duplicated, i.e. they appear both in the rigid fragment as well as in the flexible chain fragment. These are referred to as the *join atoms*. Distances of the join atoms are used to determine the compatibility of rigid fragment poses in the pose-match phase of the algorithm. The join atom positions serve the end point constraints of the flexible chain fitting, furthermore they are used to define the overlay transformation in the reconstruction of the complete solution poses before optimization.

Both the cavity and the candidate ligands are described by a Geometric Shape and Chemical Feature graph, herein referred to as GSCF graph. The nodes of the GSCF graph represent a rigid shape by a simplified geometric

hull. It is derived from regular polyhedra and then distorted to *shrink-wrap* the actual molecular fragment or cavity region (see detailed explanation of the shape generation below separately for the cavity and ligand fragment case). Chemical feature flags are associated with each vertex of the polyhedron. The edges of the GSCF graph define the connectivity between the nodes, including distance boundaries for the acceptable relative positions of the nodes.

The cavity description consists of thousands of geometric shapes (polyhedra). Center points are picked on a regular 0.5Å spacing grid, such that the point is suitable to place the center of mass of a rigid fragment. Grid cells that either violate the receptor boundary or are too close to it, are not suitable as center points. The distance from the boundary must be at least an atom radius. The space is measured in various directions from those centers and the regular polyhedra is distorted so that the vector length from the center point matches the distance measured, thus building polyhedra that represent the shape of the available space around the center. Figure 6 demonstrates the generation of a cavity node using a 2D cartoon for sake of simplicity. The 3D polyhedra overlap with each other and fill the whole cavity space. Chemical feature flags are assigned to the vertices of the polyhedra.

The distance measurement from the center to the receptor boundary is performed using a 3D steric grid, which is generated within a bounding box of the binding site. This bounding box also acts as an artificial closing of any binding pocket that is open to the solvent water. If no receptor boundary is hit by the scanning ray that is measuring the empty space in the direction of a vector, the the bounding box terminates the ray placing a practical limit on the polyhedron vector length.

The rigid fragments are also wrapped into polyhedra described by directional vectors from their centers of mass. Again the vectors from the center to the vertices of the polyhedra are scaled to match the distance from the center of

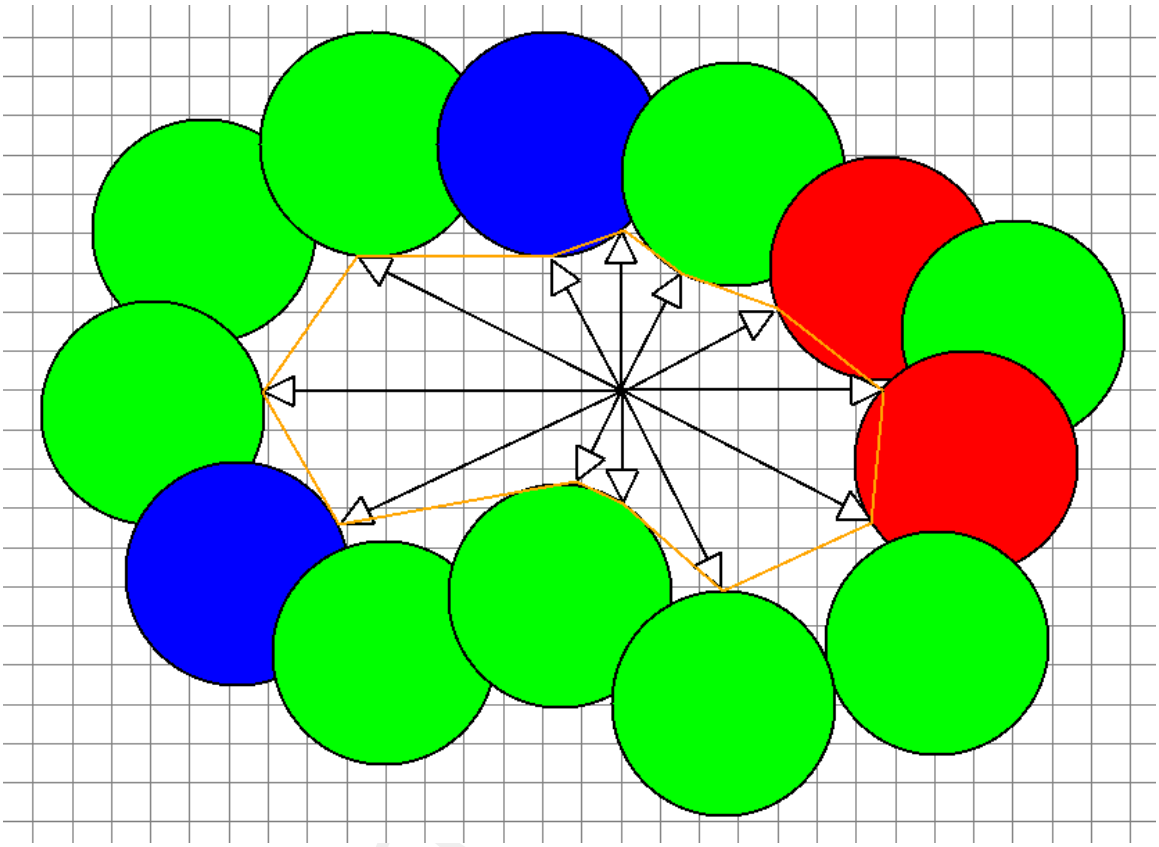


Figure 6: Simplified 2D cartoon demonstrating the generation of a cavity descriptor polygon using 12 vectors in 30 degree increments. The available space around grid points is measured in directions dictated by regular polyhedra shapes, then chemical property flags are assigned to the end points based on the chemical activity of the closest receptor atoms.

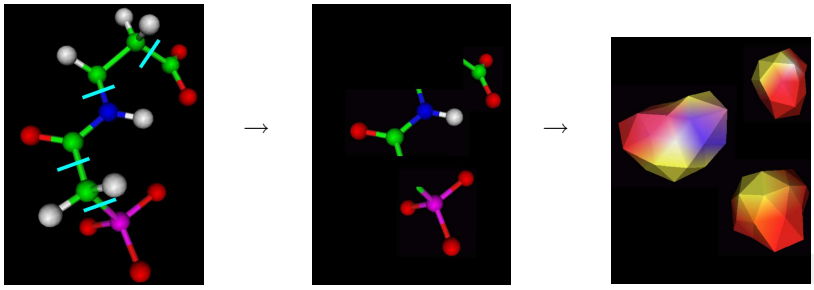


Figure 7: The ligand is broken into rigid fragments and each fragment is wrapped into a polyhedron shape with chemical properties assigned to the vertices of the polyhedron. The colors on the left and middle picture are atom type based. On the right side picture with the polyhedron shape, the colors represent the interaction flags assigned to vertices: red is hydrogen bond acceptor, blue is hydrogen bond donor, yellow is join point of the fragment.

mass of the ligand fragment to the van der Waals surface in the direction of the vector. Figure 7 shows an example of how a ligand is divided into rigid fragments and shrink-wrapped into a polyhedron shape. The polyhedron is color coded to represent the chemical features assigned to the vertices.

The polyhedrons are created by shrinking the vector lengths from the center to the vertices, but the directions are maintained, therefore the angles between them are not changed either. Consequently, if the self symmetric transformations of the regular polyhedra are applied to these polyhedra, then each directional vector from center-to-vertex will be overlayed on another such vector by the transformation. Each transformation can be described as a specific permutation of the vertices.

3.2 Rigid fragment docking

The rigid fragment docking proceeds by placing the rigid fragment polyhedra inside the cavity polyhedra. All combinations are explored (each rigid fragment polyhedron with each cavity polyhedron) and all orientations of the polyhedra. We use directional vectors based on the vertices of an icosahedron and a dodecahedron combined. These regular polyhedra have 60 self-symmetric transformations each, so we use those to orient the rigid fragment polyhedra inside the cavity polyhedra.

The polyhedron representation allows a very rapid enumeration of all fitting poses using the following method. The GSCF graph nodes contain the length of the directional vectors to each vertex, and they also contain the decreasing order of these lengths.

Step 1. The lengths of the ligand node vectors are checked against the cavity vector lengths in decreasing order. If any ligand vector is larger than its corresponding cavity vector plus ϵ grid-tolerance, then it is impossible to fit the rigid fragment node into that cavity node in any orientation, therefore no detailed orientation check is necessary, so the whole loop of the following step can be skipped without any loss of solution.

Step 2. All 60 self-symmetric transformations of the regular polyhedra (dodecahedron and icosahedron) are stored in the form of a permutation table of their vertices. A loop is run to test each of the 60 orientations, using the permutation table, in each execution of the loop. The vertices of the ligand polyhedron are mapped to the vertices of the cavity polyhedron via the permutation table. The directional vector lengths are compared and the pose is rejected if the ligand vector is longer by more than ϵ grid-tolerance for any vertex.

Step 3. For each vertex map that passes the vector length based steric check, the chemical feature flags of each vertex pair are scored and summed up to give a complete chemical fit score of the given ligand fragment pose.

Step 4. The 3D coordinates are computed for the acceptable poses based on a transformation matrix that is pre-computed and stored for each row of the permutation table.

Note that in steps 1 and 2 a specific grid-tolerance value must be applied to the comparison of the vector lengths, i.e. allow the ligand vector to be longer than the corresponding cavity vector by a small amount and reject the pose *only* if the ligand vector is longer than cavity vector plus ϵ. This ϵ grid-tolerance depends on the resolution of the 3D grid that is used to generate the cavity center points (by default $a = 0.5$Å resolution is used, but it is a user adjustable parameter, higher accuracy can be reached at the expense of more CPU time if this size is reduced). The reason is that cavity graph nodes are generated at discrete locations controlled by the grid, and it is possible that if the center is shifted by a fraction of a grid cell, then a larger fragment may fit. However, this sampling error is limited by the largest possible distance of the ideal position to the grid cell corner: $\epsilon = a\sqrt{3}/2$.

All of the chemical property flags that apply are assigned to each vertex of the polyhedron, both on the cavity and the ligand fragments. A scoring matrix is defined for the flags which contains a score for each flag-to-flag interaction pair (more details on the flag based scoring are given later in the scoring section). The score of a rigid docking pose is computed by summing all the scores of any flag pairs present on matched-up vertices between cavity and ligand.

For some larger rigid fragments, the 32 vectors of the combined polyhedra will produce a surface sampling where distance between surface points is larger than the desired 0.5Å. However, this does not limit the sampling precision of the

docking, because multiple cavity polyhedra (partially overlapping each other) are used for the mapping, so there are target positions for each ligand vector with sufficient density. The cavity polyhedra are generated on a 0.5Å spacing grid with multiple orientations considered for the same center.

Typically, the program evaluates several million mappings of the rigid fragment polyhedra to cavity polyhedra. The ones that do not fit geometrically (steric violations) are rejected and the score is computed for those that do fit. Typically, there are tens of thousands of fitting poses (10-20 thousand for small pockets and large fragments, 60-100 thousand for small fragments in large cavities).

When the number of acceptable poses is too large to handle during the next (pose matching) phase, a clustering algorithm is applied to group the poses that are close to each other in RMSD metric space and a single representative is kept from each cluster. The diversity of the poses and their coverage of the cavity site is maintained during this clustering step.

This clustering step could potentially compromise the exhaustiveness of the search if the cluster representatives do not cover the pose space with sufficient resolution. The maximum number of cluster representatives is controlled by a user adjustable parameter and by default it is set to a value that achieves a fast (sub-second) PoseMatch run-time with an average separation between representatives of about 1-1.5Å RMSD. In terms of search space sampling, this means that a sampling pose is generated within $\sqrt{3}/2$ times the separation distance from any query pose (in the worst case), while the average error from the X-ray pose can be estimated to be about 0.43Å-0.65Å. This range goes slightly higher than our desired precision, but the parameter can be adjusted to achieve more precise sampling at the cost of CPU time. There is another tolerance applied during the PoseMatch phase that is computed from the actual average separation distance between the poses. That tolerance is applied to the compat-

ibility check, i.e. comparison between join point distances and connecting chain lengths. The tolerance is dependent on the actual average pose separation, so that it counters the loss of precision, allowing the selected poses to represent their whole cluster (within the radii) for the purpose of matching instead of considering strictly the particular pose. Thus the algorithm maintains the exhaustive coverage via the use of this calculated tolerance and the ability to refine the search by adjusting the control parameters of the clustering.

It is very important to keep fragment poses that do not get good scores, because even for high affinity ligands it is possible that some fragments are acting simply as spacers and are not contributing much to the binding. In fact, analysis of the X-ray complexes in the test set shows that many contain fragments that either do not make any interaction with the protein, or even make clearly repulsive interactions. Of course, the energy loss due to the "bad" interactions must be compensated by some strong attractive interactions formed by other fragments of the ligand.

All acceptable poses of the rigid fragments are computed regardless of other fragments in the ligand. Therefore, the information about the acceptable poses of a given fragment can be reused when another ligand containing the same fragment is docked to the same receptor. This situation occurs very frequently during a virtual screening study when many thousands (or even millions) of drug-like ligands are docked to a given target receptor, because such ligands often contain some typical functional groups. The DockTable extension of eHiTS makes use of the repeating fragments to speed up the screening process by using an SQL database to store all the results of the rigid fragment docking phase. An efficient hash key (canonical name) is used for indexing the database to retrieve the previous results. If no results are stored for the given rigid fragment yet, then the docking proceeds as described above in this section, then the results are deposited to the database.

It is sufficient to store the 3D transformation and the score for each pose, therefore a space efficient storage can be achieved that requires about 1MB disk space per rigid fragment for the DB (this size does not depend on the size of the fragment but it does depend on the size of the cavity). We have run experiments screening various ligand libraries against various receptor targets and observed the speed-up curve of the docking time per ligand as well as the number of fragments deposited to the database. Significant speed-up is observed during the first few hundred to few thousand ligand docking runs, but the speed tends to level out between 5 and ten thousand ligands (the speed is 2-4 fold faster at that point than docking speed without the SQL DB). The number of commonly re-used fragments is in the order of a few thousands, therefore a limit of ten thousand fragments has been implemented in the DockTable extension of eHiTS. This limits keeps the disk space requirement under 10GB per receptor regardless of the size of the ligand library docked.

3.3 Pose matching

There are several thousand alternative poses generated and scored at the rigid docking step for each rigid fragment. The next task is to select pose-sets containing a single pose for each ligand rigid fragment such that the distances between them are compatible with sizes of the flexible chains that should connect them. In addition, they must not bump into each other.

One can think of this task as mapping the ligand graph (where each node represents a rigid fragment) on to the receptor cavity graph (where each node represents a possible placement position and orientation of a ligand rigid fragment). Such graph-mapping problems are often solved by graph algorihtms operating on a hyper-graph rather than on the graphs to be mapped. The hyper-graph is a higher order graph, where nodes represent mappings between

the original graphs.

This task is solved by clique detection on the following hyper-graph. Each node of the ligand graph is represented by a set of hyper-graph nodes, one corresponds to every accepted rigid fragment pose, i.e. the nodes of the hyper-graph represent individual mappings of ligand graph nodes to a cavity graph nodes. There are edges between those node pairs where all the following conditions hold true:

a) the nodes correspond to poses of different ligand fragments,

b) there is no steric clash between the two poses, and

c) the distance between the join points of the fragments in the given poses is compatible with the length of the chain that should connect them, i.e. it is within the interval that is possible to span by the given chain.

Maximal cliques of this hyper-graph should consists of as many nodes as the number of rigid fragments in the ligand (number of nodes of the ligand graph). Each maximal clique defines a unique docking solution. By enumerating the maximal cliques we can find all distinct docking modes of the ligand in the receptor cavity.

Figure 8 shows a simple example of an adjacency bit matrix of such a hyper-graph. The matrix M can be divided into blocks representing pose combinations between the poses of two specific rigid fragments. The example matrix corresponds to a ligand that contains 4 rigid fragments, and for the sake of simplified example we assume only 8 poses for each fragment. Rows (and columns) 1 to 8 correspond to rigid fragment number 1, rows 9-16 correspond to rigid fragment number 2, etc. The stronger lines indicate the boundary between the blocks that correspond to different rigid fragments. The stars ($\star$) mark the bits that represent edges, i.e. where the column and row index corresponds to compatible

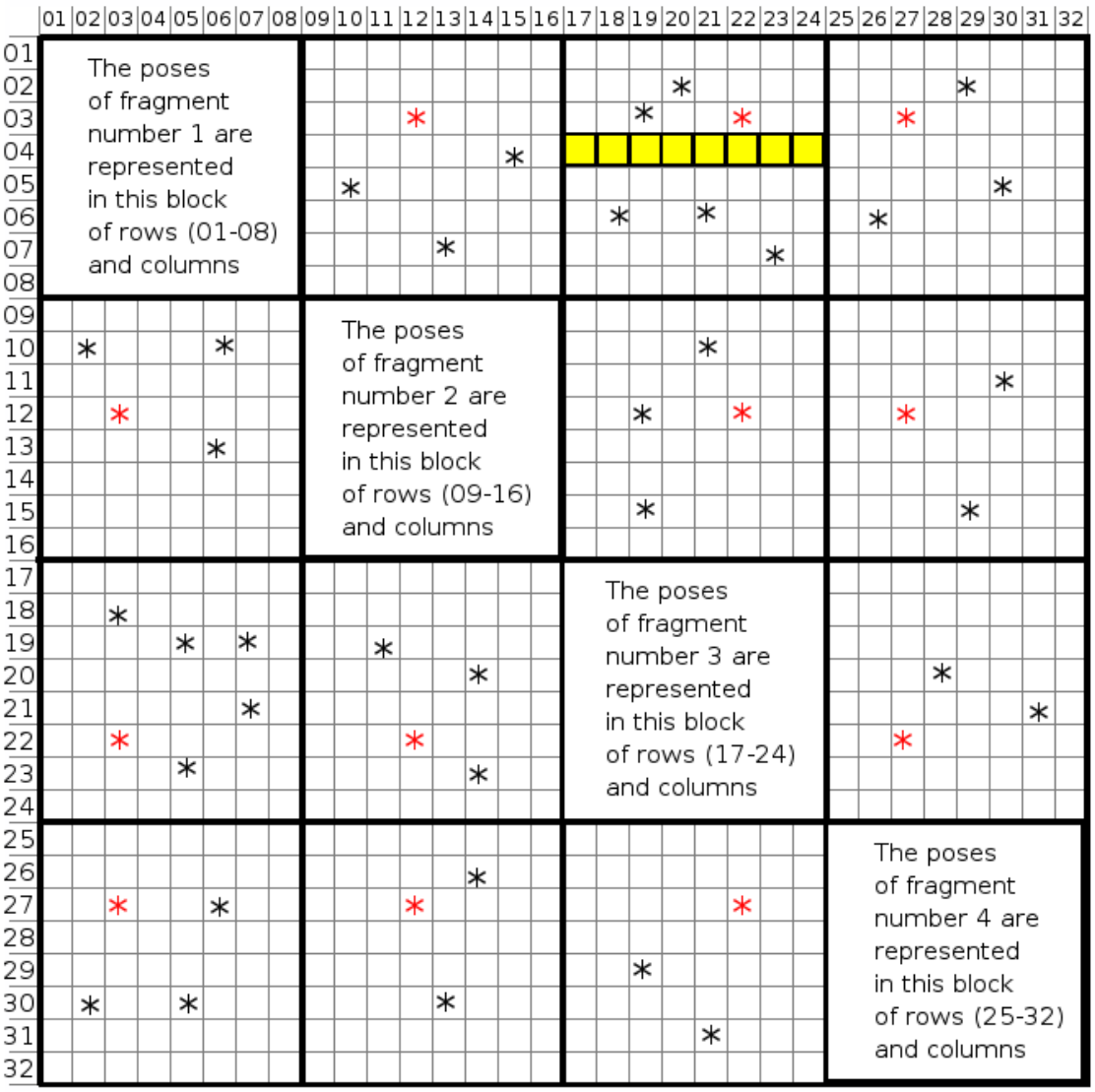


Figure 8: Example adjacency bit matrix of the hyper-graph corresponding to a ligand that consists of 4 rigid fragment. For each rigid fragment, there are 8 poses represented in the matrix. Rows and columns 1-8 correspond to poses of rigid fragment 1, 9-16 belongs to rigid fragment 2, 17-24 fragment 3, 25-32 fragment 4. The stars represent fragment pose pairs that are compatible, i.e. not bumping into each other and placed at a distance that can be spanned by the connecting chain fragments.

pose pairs. The diagonal blocks are empty, because they would correspond to alternative poses of the same node, so they are not compatible, i.e. only one pose can be selected for each node. The task is to find an S set of 4 indices such that:

$$\forall i, j \in S, i \neq j : M_{i,j} = 1$$

The red stars mark the solution maximal clique $S = \{3, 12, 22, 27\}$.

The clique detection algorithm described by Bron and Kerbrosh [3] was used as the basis for the pose matching implemented in eHiTS. The original algorithm was improved using the extra information available about the blocked nature of the adjacency bit matrix of our hyper-graph. Note, that if any row i contains an empty segment corresponding to any rigid fragment, i.e. if

$$\exists r \in \{0, ..., 3\}, \forall j \in [8r, 8(r+1) - 1] : M_{i,j} = 0,$$

then the pose corresponding to row i cannot be part of any solution, because there is no suitable pose for rigid fragment r that would be compatible with pose i. Rows 1,2,4,5,6,7 and 8 are all examples of such unusable rows (e.g. row 4 has no star in columns 17 through 24 that correspond to the third rigid fragment, this segment of row 4 is highlighted by yellow on the figure). Such rows can all be deleted to reduce the problem size before the recursive (back-track) algorithm is started. Furthermore, during the back-track algorithm, a bit row is maintained that contains the logical *and* operation of the matrix rows corresponding to the currently selected poses. If this bit-row contains an empty segment corresponding to any rigid fragment not yet represented in the clique, then it is not possible to find a completion to the current set, so the whole search tree branch can be cut and the algorithm steps back to choose a different candidate pose for an earlier rigid fragment.

With this problem specific optimization, the algorithm becomes very efficient. In fact the worst case complexity is no longer exponential as it was for

the general case, but a polynomial bound can be defined, where the degree of the polynomial is equal to the number of rigid fragments.

Each maximal clique found in the hyper-graph defines a different docking solution by selecting a pose for all the rigid fragments of the ligand in such a way that they do not bump into each other and the distances between them are compatible with the lengths of the flexible chains. The 3D coordinates of all atoms within rigid fragments are defined for every solution and the sum of the scores of the rigid fragments give a very good indication of the total interaction score that can be achieved by each solution. Even though the number of solution cliques may be large (it is several million for some examples), global scoring information is available for them at very low cost (summing up a handful of pose scores), so it is feasible to evaluate them all and select the most promising candidates for further processing.

Note, that selecting a subset of solutions at this point in the process does not compromise the exhaustiveness of the algorithm since the selection is based on global scoring information. All solutions are enumerated exhaustively, the number of PoseMatch solutions is the total number of distinct docking modes possible. The search engine must be exhaustive in order to be able to present all potential solutions to the scoring function for evaluation, as achieved here.

As explained in the scoring section, the full detailed and sensitive scoring function is not employed at this phase, but a faster, crude (greedy) function is employed. The final scoring function has also been tested in the rigid docking phase, but it was found to be inferior to the crude function in selecting the correct poses. This result can be explained by the fact that the final scoring function is too sensitive to precise interaction geometries, therefore it can only differentiate and rank optimized poses correctly.

3.4 Flexible chain fitting

Following the rigid fragment pose set selection, it becomes necessary to deal with the rotatable bonds joining them, i.e. the challenge of flexible chain fitting. However, this task is much simpler than is the case in the general flexible docking problem, because two atom positions at each end of the chain are already fixed, as they are given by the join atoms of the selected rigid fragment poses.

The task is to find a dihedral angle sequence that will lead from the given starting points to the given end points while avoiding steric clashes with the receptor boundary and the rigid fragments along the way. For smaller chain lengths, even analytical calculation of the complete algebraic solution space would be feasible without considering the steric boundary conditions.

A more general approach has been chosen to find a suitable set of dihedral angles that bridge the distance between the atom pairs and avoids steric clash with the receptor while preferring angles near low energy rotomers. First, a lookup table is used to select initial candidate chain conformers that consist of low energy dihedrals that have ending atom pair distances similar to those required. Then a local minimization is performed to tweak the dihedrals to reach the exact required distances.

For the lookup table, a double diamond lattice is used, which contains all pathways consisting of staggered and gauche dihedrals up to the desired number of bonds. The lattice is positioned on the starting atom pair, then the ending atom pair positions are used to locate nearby atoms in the lattice. The lookup table associated with the diamond lattice contains information about the path lengths (number of bonds from the starting atom) for each atom of the lattice. Any path with the required number of bond that ends within 3Å of the desired 3D coordinates will be considered. A deterministic minimization, based on the partial least squares fit method, is applied to tweak the chain until the end

points match precisely and no severe boundary violations occur. This tweaking method may produce *any* dihedral necessary to reach the end points and resolve clashes – even the highest local energy eclipsed conformation is allowed, if necessary. However, the local optimization starts out with low energy rotomers and will only apply the minimum necessary distortion to resolve steric clashes and bring the end points closer to the goal, so the tweaking process stops with a chain conformation with the lowest energy dihedrals that are suitable for the requirements.

There is no discrete sampling applied in this dihedral refinement process, the precision is only limited by the floating point representation of the computer. Therefore the dihedral angle sampling of eHiTS is practically equivalent to continuous (infinitesimally small) sampling.

3.5 Reconstruction and optimization

When all the flexible chains have been fitted to the rigid fragment poses, the complete ligand is reconstructed from the fragments.

Each hyper-graph clique defines a separate solution. Each solution is constructed by pair-wise joining of the rigid fragments in the selected pose with the flexible chains fitted to them. The mapped pose of each rigid fragment and the resulting conformation of the flexible chain fitting are overlayed using the two atoms that form the broken bond. These two atoms were replicated in both the rigid fragment and the flexible chain, so they can be used to drive the reconstruction.

The flexible chain fitting minimization process attempts to position the last two atoms of the chain to overlay with the target rigid fragment, however, it is not guaranteed that perfect (zero distance) match can be achieved. In other words, the join atoms on the rigid fragments and those on the flexible chain

may have different coordinates. Small transformations are carried out on the fragments to achieve complete overlay of the join atoms prior to reassembly of the complete ligand. This step ensures that all bond lengths and angles are maintained from the input structure.

A continuous local energy minimization which only allows torsional changes and rigid body transformations (rotations and translations) is applied to the complete ligand to refine binding geometries and resolve any sampling roughness from the initial polyhedron based rigid fragment positioning. A steepest descent downhill optimization is applied on $6 + n$ variables (where n is the number of rotatable bonds) to improve the scoring function value using the modified Powell's algorithm [21]. The free variables of the optimization correspond to 3 degrees of translation, 3 degrees of rigid body rotation and n degrees of torsional conformation freedom.

The precision of atom positions obtained in this phase are not limited to any discrete sampling, they are again limited only by the precision of the floating point representation of the computer. The optimization is terminated when the scoring function value does not improve in any direction in the $6+n$ dimensional transformation space, i.e. local minimum is reached.

The objective function includes interaction scoring components between the receptor and the ligand, as well as internal intra-molecular interaction components within the ligand and conformational strain energy for the sub-optimal dihedral angles. As a result, eHiTS is capable of generating strained dihedral angles, where necessary, when compensated by the interaction energy - as observed in many experimental crystal structures. However, the program will prefer the low energy conformers when they are suitable for the docking pose.

There is no stochastic element in this applied optimization technique, because the goal is to find a *local* minimum of the objective function for every particular solution. The global coverage of the search space is guaranteed by

the full cavity coverage of the rigid fragment docking step and the exhaustive algorithm of the pose matching step.

3.6 Protonation handling

The issue of protonation state is very important to the docking problem. Ligands and receptors with different protonation states can have dramatically different binding poses. However, it is common practice for many docking programs to ignore this issue and require that the user define a particular protonation state prior to running a docking experiment.

Protonation states of ligands and receptors are determined by the interaction between the two.Thus for any particular receptor-ligand pair there will generally be one correct protonation state. However for a different ligand, the protonation state of the receptor may be altered, to reflect the characteristics of the ligand. If a docking program were to pre-set the protonation state of the receptor then possible interactions with a ligand could be lost.Similarly, presetting the protonation states of ligands in a library would produce incorrect results with respect to certain receptors. A better solution, with a more appropriate score, can be found only if the program is run with various protonation states (not necessarily the neutral or the normally lowest energy form of the receptor or ligand on its own or in solvent, but the form required to reach the lowest energy for the complex).

The molecule in Figure 9 has 150 possible protonation states. Table 2 shows the 5 possible protonation states for each of A and D, 2 for B and 3 for C, combined this leads to $5*5*2*3 = 150$ different possible protonation states. Although, two pairs of states for A and D can be considered equivalent via rotations about the bond to R (swapping the roles of the 2 oxygen atoms), so a flexible docking program could work using only 3 protonation states for those

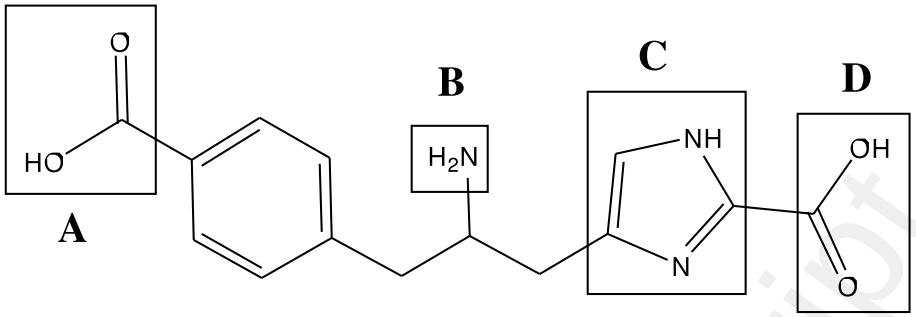


Figure 9: Sample ligand with 150 different possible protonation states. Functional Groups A, B, C and D each have multiple forms depending on protonation.

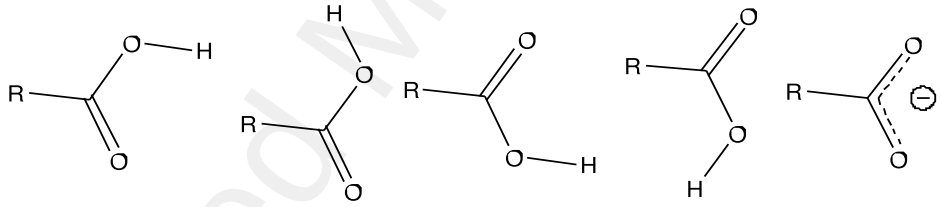
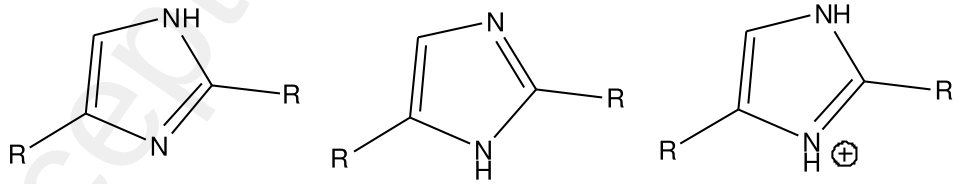
Group	Protonation states
A, D	
B	$R-NH_2, R-NH_3^+$
C	

Table 2: Protonation states of the Functional Groups from Figure 9

fragments giving a total of 3*3*2*3=54 instead of 150. Most docking programs would need to dock all 150 (or at least 54) combinations separately to evaluate the different possibilities, not even considering different protonation states of the receptor.

eHiTS takes a unique approach to the protonation problem by systematically evaluating all possible protonation states for both the receptor and ligand efficiently in a single run. Ambiguous properties flags are assigned for positions that could be either protonated or deprotonated (i.e. have a lone pair).Then during the docking algorithm both states of such surface points are evaluated and scored, selecting the best protonation state for each individual interaction independently, thus avoiding the combinatorial effect of multiple functional groups with variable protonation states. The results of a single eHiTS run using the ambiguous properties flags contain the cummulative results that would be achieved by running many individual docking runs with fixed protonation states considering all ligand protonation states (150 in the above example) against all receptor protonation states (usually an even higher number).

4 Scoring function

There are three different scoring functions used in the eHiTS process. First, a simple and fast chemical flag based statistical scoring function (SF_s) is used during the rigid fragment docking and pose matching phases. This function is not too sensitive to small variations in the interaction geometry, interaction distance and hydrogen bonding angle.

A more sensitive, empirical scoring function (SF_e) is used during the final local energy minimization phase. This scoring function has smooth curves representing the distance and angle dependency of the interactions while supporting efficient gradient based optimization.

The final result poses are evaluated by a third, more time consuming scoring function (SF_c) that combines both statistical and empirical components, plus additional grid based geometrical terms as well as entropy loss estimation and another novel scoring element based on the coverage of receptor surface area. This final scoring function attempts to estimate the binding free energy more accurately. The result of the final accurate scoring is used to rank the generated solutions.

4.1 Surface map based statistical score (SF_s)

We score the receptor-ligand interactions based on the molecular surface contacts that occur using a fine (0.5Å) resolution surface point sampling. Chemical property flags are assigned to each surface point both on the ligand and on the receptor. The receptor surface points also receive an assigned weight based on the pocket depth in the cavity, i.e. deeper points receive higher weights than shallow points. A flag compatibility matrix is used to assign scores to interacting point pairs and the point-scores are weighted by the pocket depth. Exposed receptor points (where no ligand point is within interaction distance) are assigned a penalty score based on the depth value. The sum of all receptor surface point scores is computed.

The list of chemical property flags includes:

- Strong hydrogen bond donor: D
- Weak hydrogen bond donor: d
- Strong hydrogen bond acceptor: A
- Weak hydrogen bond acceptor: a
- Strong hydrophobic/lipophilic atom: H
- Weak hydrophobic/lipophilic atom: h

- Aromatic, π stacking ring atom: R
- Aromatic, π stacking ring center: O
- Aromatic, π stacking ring edge: E
- Partial negative charge: n
- Partial positive charge: p
- Formal negative charge: N
- Formal positive charge: P
- Metal ion with unoccupied coordination: M
- Ligand atom suitable to bind to a metal ion: m

The feature flags used are listed above. An example of their use is shown in Figure 10 demonstrating the way these flags are assigned to a Histidine side-chain. There are strong and weak hydrogen bonding and hydrophobicity flags, aromatic Pi-stacking indicator flags, partial and formal charge indicators and metal ion indicator (M) with corresponding ligand flag (m) indicating which atoms can bind to a metal ion. All applicable flags are assigned to a point as the figure illustrates. Note, that alternative protonation states are handled by multiple flag assignment as shown on the figure by the red highlight: both Nitrogen atoms have the same flag-set towards the edge of the ring DApnmE, which indicates that it may be a H-bond donor or acceptor with either positive or negative partial charge and that it can interact with a metal ion or participate in aromatic Pi stacking as a ring-edge. This assignment simultaneously handles all three protonation states of Histidine.

One may notice the use of weak secondary interaction flags. Experimental data proves that some polarized carbon atoms (e.g. in an aromatic ring next to

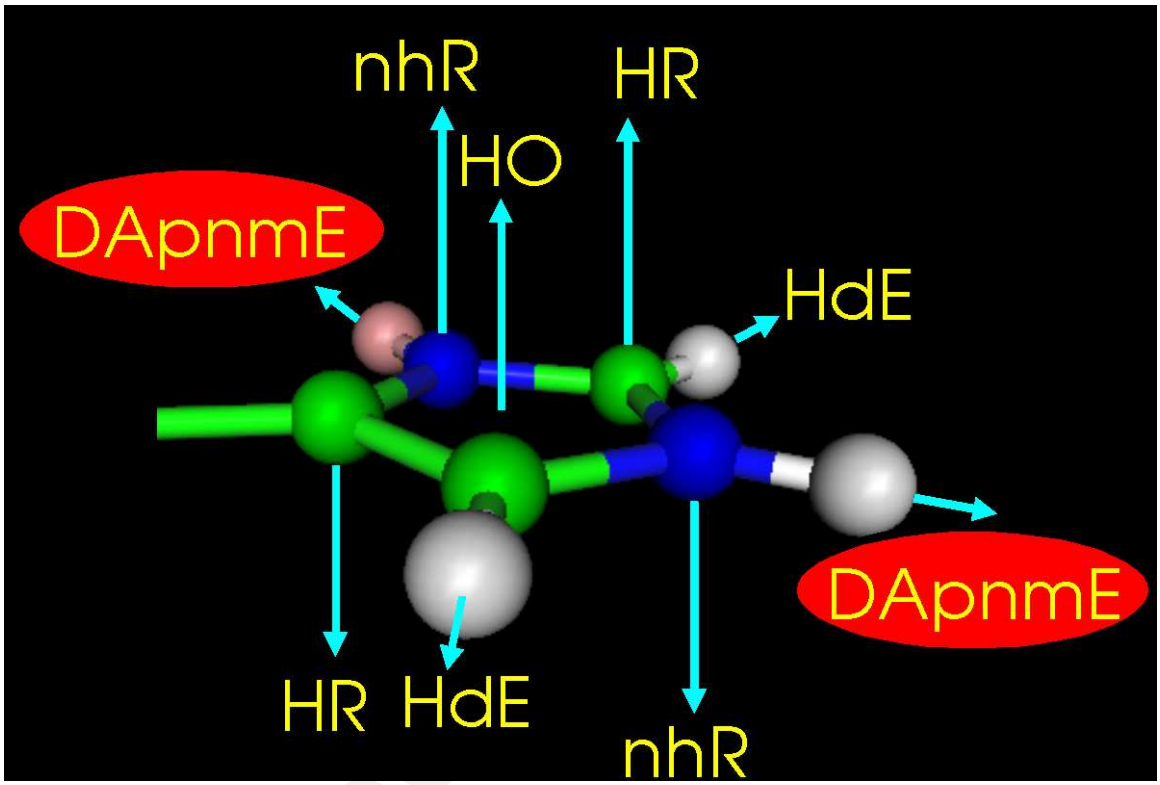


Figure 10: Example of the assignment of chemical property flags for the ring of the Histidine residue

a nitrogen atom) can act as hydrogen bond donors[5]. On the other hand, the heteroatoms of an aromatic ring are given secondary hydrophobic flags above and below the plane of the ring. This fine property assignment will correctly distinguish between the different electrostatic and chemical properties expressed from the same atom in different directions, i.e. towards the edge and the face of the ring. This is a major advantage of this surface based scoring system in comparison with the traditional atom based scoring.

4.2 Empirical scoring function (SF_e)

During the optimization phase of the eHiTS process, an empirical scoring function is employed which has the following terms:

- Hydrogen bonding term (D, H, A and Lp represent the positions of donor, Hydrogen acceptor atoms, and lone electron pair respectively):

$$E_{H-bond} = E_{max} * f_{dist}(\|H \rightarrow A\|) * \cos(\angle(A \rightarrow Lp, D \rightarrow H))$$

- Hydrophobicity term (S is the set of surface points with flags H or h):

$$E_{Lipo} = \sum_{p \in S_{lig}} E_{max} * f_{dist}(p, q_p), \quad q_p \in S_{rec} : d(p, q_p) = \min_{q \in S_{rec}} d(p, q)$$

- Aromatic π stacking (similar to E_{Lipo}, but applied on surface points with flags E,R and O with different E_{max} for each type of pair)
- Electrostatic potential (Coulomb formula)
- Van der Waals contact energy (Lennard–Jones formula)
- Metal ion interactions (distance and angle dependency similar to the Hydrogen bonding term)

- Penalty for incompatible contacts (e.g. polar-hydrophobic, same charge, evaluated similarly to E_{Lipo}, but for surface points with incompatible flags)
- Exposed surface atoms are scored against solvent properties
- Intra-molecular interactions are taken into account: ligand (conformation score) and receptor (protonation, competing interactions reduce the score towards ligand interactions)

Each term is configurable via a text parameter file. The base functions for distance dependency (f_{dist}) can be selected from a predefined set of functions. The coefficients (E_{max}) and weights of the terms can be set or changed from their defaults to fine tune the function.

4.3 Final ranking score (SF_c)

For the final ranking of the generated solutions, the empirical scoring terms are all computed. A statistical term is computed similar to the one described earlier with the exception that this time each surface point pair score is multiplied by a contact geometry factor, which is determined by the distance of the surface points from each other and the angle of the surface normal vectors. Statistical distribution of these contact geometry descriptors have been collected from receptor-ligand interactions in 5000 high resolution X-ray crystal structures. The weighting factor for each surface point pair represents the fit of their actual contact geometry to the statistical histogram of distance and angle distribution.

The following terms are also added to the final score:

- Total surface contact area between receptor and ligand
- Hydrophobic surface area of the receptor that is *not* buried by the ligand

- Exposed hydrophobic surface area of the ligand
- Sum of pocket depth values for all ligand atoms

All empirical terms, the statistical score and the above listed extra terms are combined with an adjustable set of weights to form the final score value. The weights have been calibrated using 133 receptor ligand complexes for which both high resolution X-ray crystal structures (in PDB) and corresponding experimental binding energy values were publicly available. A fully detailed discussion of this scoring function (including the method of calibration) will be published elsewhere.

5 Results

In choosing a validation set to illustrate the docking accuracy and ranking capability of the eHiTS process, it was decided to select a set of protein complexes previously reported by other docking software. Gold and FlexX are widely considered among the most popular docking programs and Glide has been very active in the market place over the past couple years. In examining the performance of eHiTS, comparisons were made with these three docking programs.

Validation results have been reported for the above mentioned docking programs allowing us to report a head-to-head comparison. Validation results on 282 protein ligand complexes were reported for Glide[8], 306 for GOLD[12], and 200 for FlexX[22]. The results presented here are from the 91 complexes common to all three test sets, for which individual rms values were reported[1].

[1] 1aaq 1abe 1acj 1acm 1aco 1aha 1apt 1azm 1baf 1cbx 1coy 1cps 1dbb 1dbj 1did 1die 1dr1 1dwd 1eap 1eed 1epb 1eta 1etr 1fkg 1fki 1ghb 1glq 1hdc 1hef 1hri 1hsl 1hyt 1icn 1ida 1igj 1ive 1ldm 1lic 1lst 1mcr 1mdr 1mrk 1mup 1nis 1pbd 1phd 1phg 1poc 1rds 1rne 1rob 1slt 1srj 1stp 1tdb 1tka 1tmn 1tpp 1ulb 1xid 1xie 2ack 2ada 2ak3 2cgr 2cht 2ctc 2dbl 2mcp 2phh 2pk4 2plv 2r07 2sim 2yhx 3cla 3cpa 3hvt 3mth 3ptb 3tpi 4aah 4cts 4dfr 4fab 4phv 5p2p 6abp 6rnt 7tim

This set of 91 PDB codes contains a diverse set of proteins from 66 distinct protein families. For all results shown, the standard (default) parameter set of eHiTS was used. This set results in docking times of a few minutes/ligand on a standard desktop PC.

It is important to note that no manual preprocessing was performed on any of the selected PDB complexes. The protonation states, cofactors, counter-ions, solvent molecules, partial charge assignment, etc. were all handled by eHiTS without user intervention. This automation makes eHiTS very user-friendly and capable of automated processing.

The ligands were docked into the original protein binding site (as provided in the X-ray structure) and the accuracy was measured by calculating the root-mean-squared deviation (RMSD) between the coordinates of the heavy atoms of the ligand in the eHiTS docked pose and those in the crystal structure. Statistics on RMSD values of the highest ranked pose by the eHiTS scoring function are reported.

The result would further improve if eHiTS was allowed to spend more time by local minimization of larger number of partial solutions produced by the pose matching step. The default parameters set allows 200 poses to be entered into the optimization phase, while the accurate parameter set would allow 640 solutions, reducing the chance of dropping the correct pose based on selection by the initial rough scoring function. The program often generates millions of distinct solution poses in the pose matching phase (all of them fit the steric constraints and exhibit attractive chemical interactions), therefore processing all of them in the energy minimization phase would be impractical. The search engine has proven to be exhaustive within the desired precision, i.e. the closest pose produced at the early stages of the search engine pipeline is found to be under 1Å RMSD for all test cases, with an average of 0.4Å RMSD.

8gch

The poses selected for minimization and finally output do not reach that precision, indicating that there is still room for improvement concerning the scoring function. All three components of the scoring function are responsible for some loss of precision. Score based selection in PoseMatch often rejects the solution with the smallest RMSD. The final local minimization phase has a difficult task, because the scoring function has a too volatile (chaotic) shape with many local minima close to each other in the $(6+n)$ dimensional transformation space. If the scoring function could be made smoother with a funnel shaped local minimum near the X-ray pose that would significantly improve the accuracy of the optimization step. Finally, the ranking score often does not pick up the closest solution to be top ranking.

5.1 Comparison to GOLD

GOLD is a stochastic program that employs a Lamarkian Genetic Algorithm, encoding the ligand conformation and the mapping of interactions points between ligand and receptor atoms. The program employs an island model, creating several small populations rather than one large one. The genetic operations include migration of individuals from one population to another, crossover and mutation. The fitness of a new individual is assessed using 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.

EHiTS contains no stochastic elements, ensuring a systematic coverage of the search space, thus the two algorithms are fundamentally different.

Table 3 shows a summary of the docking results for eHiTS, GOLD and FlexX. On average eHiTS gives rms deviations less than half of those given by GOLD. Percentages of cases docked under specific limits of 0.5Å,...,2.5Å are

also favorable for eHiTS.

5.2 Comparison to FlexX

FlexX employs an incremental construction algorithm, in which a seed fragment (base) is identified first and placed into the active site. Subsequently a tree search is performed that adds fragments (using a discrete set of dihedral angles) incrementally and evaluates the intermediate substructures, eliminating those which score badly. EHiTS differs significantly from incremental construction. Firstly, there is no bias set by selecting an initial base fragment, since all rigid fragments are docking independently. Secondly, eHiTS makes global decisions on whether or not to keep a particular ligand pose, thus allowing for bad partial structures, if the global score is good. Thirdly, eHiTS does not use a discrete set of torsions, when rigid fragments are connected a tweaking algorithm permits any torsion, limited only by the floating point precision of the computer.

As Table 3 shows, on average the top ranking solution of eHiTS gives rms deviations less than 40% of those given by FlexX. Furthermore, the eHiTS solutions fall within desired precision limits (e.g. 1.0Å, 1.5Å or 2.0Å) about twice as often as the FlexX solutions.

5.3 Comparison to Glide

The Glide algorithm combines a systematic pose search with a stochastic refinement and optimization procedure. A relatively small number of low energy conformers are generated for the core of the ligand containing most of the rotatable bonds after the removal of terminal rotomer groups. These conformers are systematically positioned on a 2Å grid and evaluated with a discretized ChemScore scoring function (using precomputed scores from a 1Å spacing grid). A small number of best scoring poses are further refined and optimized using a

stochastic technique (Monte Carlo Simulated Annealing) with the OPLS-AA force field energy function. In comparison, eHiTS does not rely on a discrete set of low energy conformers of the flexible components of the ligand, nor does it rely on interactions of a core fragment for filtering. All rigid fragments of the ligand are treated equally in eHiTS and positioned exhaustively on a much finer 0.5Å grid (considering the cubic effect of the 3 dimensional space, a 0.5Å grid yields 64 times finer sampling than a 2Å grid), and evaluated with a scoring function that is discretised with 0.2Å binning, which yields 125 times more precise representation in 3 dimensions than the 1Å spacing scoring grid of Glide. In the final minimization phase, eHiTS uses a deterministic gradient optimization avoiding the use of stochastic methods, because, as correctly stated in the Glide paper[8]: "such methods can miss key phase-space regions a certain fraction of the time, thus precluding development of a truly robust algorithm".

Direct comparison of the rms deviation values reported for Glide is unfair to other docking programs as acknowledged on page 1743 of the cited publication[8] due to the fact that the calculation was not made relative to the X-ray pose. Glide requires a preprocessing step, where the protein-ligand complex is minimized by the OPLS-AA force field energy function (same as the final scoring function that drives the minimization in Glide) to "anneal away" steric clashes, orient the hydroxyl groups, determine protonation states and slightly alter the atom positions of both ligand and receptor to optimize the local energy according to the force field. The input to Glide is the altered receptor file and its output result rmsd values are computed relative to the ligand pose altered by the preprocessing optimization. Considering that the same scoring function is used to minimize the poses at the end of the Glide run, from a mathematical standpoint, the calculated rms deviation is a property of the OPLS-AA force field energy function, i.e. the distance of two specific local minima, one close to the X-ray pose and another one close to the pose generated by Glide. This fact

is further supported by the observation reported by E. Perola *et al*[20]: when the top 20 poses generated by GOLD are subjected to the same minimization on the OPLS-AA force field, the results become equivalent to those generated by Glide.

The mathematical meaning of such RMSD calculation is further demonstrated on a simplified example using a hypothetical 1D scoring function on Figure 11. On the figure, point X represents the X-ray pose of the ligand, point O is the pre-optimized pose (local minimum of the scoring function corresponding to the well of X). Suppose, the docking program has generated raw poses R and P during its crude positioning phase. During local minimization they are transformed into poses T and C respectively (by downhill minimization on the scoring function). The correct RMSD measurement would be the distance $\|T - X\|$ for top-ranking and $\|C - X\|$ for closest solution. Using the pre-optimized pose O, yields distances $\|T - O\|$ and $\|C - O\|$ instead. These are distances between local minima of the scoring function (i.e. a mathematical property of the scoring function shape) and insensitive to the exact position of X within the shaded box corresponding to the region $[A, B]$. If there is any raw docking pose generated anywhere within the region $[A, B]$, then the RMSD of the final (optimized) output pose from O would be zero. Therefore, calculating RMSD against a pre-optimized pose is not a suitable measure of docking pose accuracy. An idealistic scoring function would have a perfect funnel shape with a single local minimum, in which case *any* raw docking pose against *any* pre-optimized X-ray pose would yield zero RMSD, clearly demonstrating problems with such measurement.

Similar preprocessing optimization can be performed with any scoring function. The choice of scoring function for the preprocessing will strongly bias the results in favor of the docking program that employs the same scoring function in its final optimization phase. To demonstrate the effect of such pre-optimization

Program name	Average rms	cases rms<0.5Å	cases rms<1Å	cases rms<1.5Å	cases rms<2Å	cases rms<2.5Å
eHiTS	1.47Å	15%	42%	67%	77%	85%
GOLD	3.11Å	8%	35%	55%	65%	68%
FlexX	3.87Å	3%	20%	37%	49%	56%

Table 3: Comparison of docking accuracy of three programs on a set of 91 protein-ligand complexes. The rms deviations are computed between the top ranking solution and the X-ray pose in the original PDB structure.

Program name	Average rms	cases rms<0.5Å	cases rms<1Å	cases rms<1.5Å	cases rms<2Å	cases rms<2.5Å
eHiTS	0.66Å	45%	88%	98%	100%	100%
Glide	1.87Å	29%	46%	62%	71%	77%

Table 4: Comparison of docking results of eHiTS and Glide on a set of 91 protein-ligand complexes. The rms deviations are computed between the top ranking solution and the pose obtained by local minimization of the X-ray pose with the scoring function of the program.

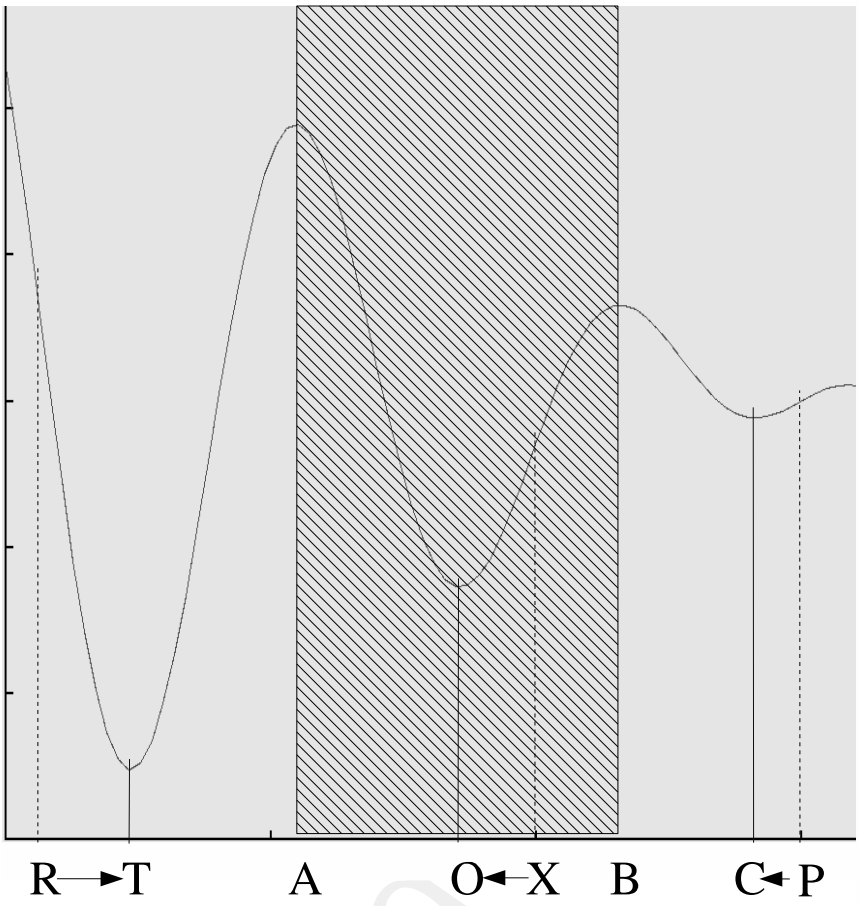


Figure 11: Hypothetical example demonstrating the mathematical meaning of measuring RMSD against pre-optimized pose instead of the X-ray pose. Point X represents the X-ray pose of the ligand, point O is the optimized X-ray pose, T is the top-ranking solution pose (derived from R raw docking pose with optimization), C is the closest solution pose (derived from P raw docking pose with optimization).

and result rms calculation relative to the optimized pose instead of the original X-ray pose, an equivalent preprocessing of the X-ray ligand was performed using the eHiTS scoring function and then the result rms values were re-computed relative to the optimized pose. The results of this type of RMS calculation are shown in Table 4 in comparison to the rms values reported for Glide. Note, that the same eHiTS result poses were used as in Table 3, only the RMS values are smaller here because the target of the comparison is changed from the X-ray pose to the optimized pose.

It should be emphasized, that these values represent a property of the scoring function employed, i.e. distance of two specific local minima instead of the true accuracy of the search engine employed. Since different scoring functions are employed by eHiTS and Glide, the comparison of these results have little meaning. Nevertheless, results generated by eHiTS are superior to the results generated by Glide even if the rms deviations from the X-ray structure (in Table 3) are considered for eHiTS.

6 Conclusion

The new algorithm employed in eHiTS permits an exhaustive flexible docking with a very fine sampling of atom positions for rigid fragments (less than 0.5Å error guaranteed) and a continuous tweaking of dihedral angles for rotatable bonds. Complete processing of all partial results can lead to the generation of millions of sterically fitting binding poses for test cases with a large receptor cavity and a ligand with many small rigid fragments. The eHiTS run can be controlled with user adjustable parameters to select the desired compromise between docking accuracy (closest solution under 1Å RMSD for most cases in highest accuracy mode) and execution time (a few seconds per ligand for fastest runs). The default parameter set provides a balance where the execution time

is in the range of a few CPU minutes per ligand per processor, and solutions within 2Å RMSD from the X-ray pose are found for most of the test cases. Alternative parameter sets are also provided to achieve more accurate results or faster run times.

To summarize the advantages of eHiTS, it offers a truly exhaustive search method, which is important to minimize the chance of false negatives during a virtual screening study. eHiTS offers highly accurate (energy minimized) docking poses. It is a deterministic system, so the results are completely reproducible – an important feature for the scientific method.

eHiTS is a fully automated system, which does *not* require lengthy data preparation. There is no need to manually assign protonation states, partial charges, add hydrogen atoms or lone electron pairs, since eHiTS does that all automatically. It can even detect automatically the location of the binding pocket if the full protein is supplied without any further information. The scoring function is highly configurable via simple text parameter files. And last, but not least eHiTS can run very fast using the DockTable extension that reuses fragment docking information for common functional groups in large databases or virtual libraries.

The eHiTS software is available from SimBioSys, Inc. (www.simbiosys.ca) for interested researchers on Linux and IRIX platforms. It is free of charge for Academic use. The results reported here were obtained running the software on Pentium4 processor with 1GB RAM under Linux, but the typical memory footprint of an eHiTS run is in the range of 50MB-300MB depending on the size of the receptor cavity and the ligand structure. The software requires about 2MB to 300MB disk space for the preprocessing data (grid and cavity graph) of each receptor. The size of the result files varies with the size of the ligand and the number of poses generated, typically it is in the range of few MBs.

7 Acknowledgement

We would like to thank the National Research Council and the Government of Canada for financial support of our research work under IRAP projects #411461 and #468906, and SR&ED projects in fiscal years 2001 and 2002.

We would like to thank former and current colleagues, Irina Szabo, Zsolt Szabo, David Fung, Sing Yoong Khew, James Law, Constantin Tanurkov, and Beihong Wu for their contributions in the implementation and testing during the eHiTS software development.

We would like to thank the scientists at pharmaceutical companies for their valuable comments and suggestions made in the course of testing eHiTS, including Istvan Enyedi (Bayer), Tanja Schulz-Gasch (Roche), Chaya Duraiswami (GSK) and Bruno Bienfait (NuadaPharma).

References

[1] J. Bostrom, Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools, . *J. Comput.-Aided Mol. Des.* 15, **2001**, 1137–1152.

[2] J. Bostrom, Norrby, Per-Ola, and T. Liljefors, Conformational energy penalties of protein-bound ligands, *J. Comput.-Aided Mol. Des.* 12 **1998**, 383–396.

[3] Coen Bron and Jeep Kerbosch, Finding all cliques of an undirected graph [h], *Comm. ACM* 16, **1973**, no. 9, 575–577.

[4] Cambridge Crystallographic Data Centre, Cambridge, England, Cambridge structural database system user's manual, 1989.

[5] Z. S. Derewenda, U. Derewenda, and P. M. Kobos, (his)ce-h...o=c¡ hydrogen bond in the active sites of serine hydrolases, *J. Mol. Biol.* 241 **1994**, 83–93.

[6] R. L. DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, and I. D. Kuntz, Using shape complementarity as an initial screen in designing ligands for receptor binding site of known three-dimensional structure, *J. Med. Chem.* 31, **1988**, no. 4, 722–729.

[7] J. L. Doob, The brownian movement and stochastic equations, *Ann. Math.* 43, **1942**, 352–369.

[8] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Kicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter Shenkini, A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy, *J. Med. Chem.* 47, **2004**, 1739–1749.

[9] D.S. Goodsell and A.J. Olson, Automated docking of substrates to proteins by simulated annealing, *Proteins: Structure, Function, and Genetics* 8, **1990**, 195–202.

[10] T. N. Hart, S. R. Ness, and R. J. Read, Critical evaluation of the research docking program for the casp2 challenge, *Proteins Suppl.* 1 **1997**, 205–209.

[11] G.A. Jeffrey and W. Saenger, Hydrogen bonding in biological structures, Springer Verlag, Heidelberg, 1991.

[12] G. Jones, P. Willett, R.C. Glen, A. R. Leach, and R. Taylor, Development and validation of a genetic algorithm to flexible docking, *J. Mol. Biol.* 267, **1997**, 727–748.

[13] S. K. Kearsly, D. J. Underwood, R. P. Sheridan, and M. D. Miller, Flexibase: a way to enhance the use of molecular docking methods, *J. Comput. Aided Mol. Des.* 8, **1994**, 565–582.

[14] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications, *Nat Rev Drug Discov.* 3, **2000**, no. 11, 935–949.

[15] Ming Liu and Shaomeng Wang, Mcdock: A monte carlo simulation approach to the molecular docking problem, *J.Comp.-Aided Mol. Des.* **1999**, 435–451.

[16] M. McGann, H. Almond, A. Nicholls, J.A. Grant, and F. Brown, Gaussian docking functions, *Biopolymers* 68, **2003**, 76–90.

[17] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson, Automated docking using lamarckian genetic algorithm and an empirical binding free energy function, *J. Comp. Chem.* 19, **1998**, 1639–1662.

[18] M.C. Nicklaus, S. Wang, J.S. Driscoll, and G.W.A. Milne, Conformational changes of small molecules binding to proteins, *Bioorg. Med. Chem.* 3, **1995**, 411–428.

[19] E. Perola and P.S. Charifson, Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding, *J. Med. Chem* 47, **2004**, 2499–2510.

[20] Emanuele Perola, W. Patrick Walters, and Paul S. Charifson, A detailed comparison of current docking abd scoring methods on systems of pharmaceutical relevance, *Proteins: Structure, Function, and Genetics* 56, **2004**, 235–249.

[21] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, Numerical recipies in c++, Cambridge University Press, Cambridge, UK., 2002.

[22] M. Rarey, B. Kramer, T.Lengauer, and G. A. Klebe, A fast flexible docking method using an incrememental construction algorithm, *J. Mol. Biol.* 261, **1996**, 470–89.

[23] William Welch, Jim Ruppert, and Ajay N Jain, Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites, *Chemistry and Biology* **1996**, 449–462.

[24] D. R. Westhead, D. E. Clark, and C. W. Murray, A comparison of heuristic search algorithms for molecular docking, *J. Comput. Aided Mol. Des.* 11, **1997**, 209–228.

8 Table of Contents Graphic

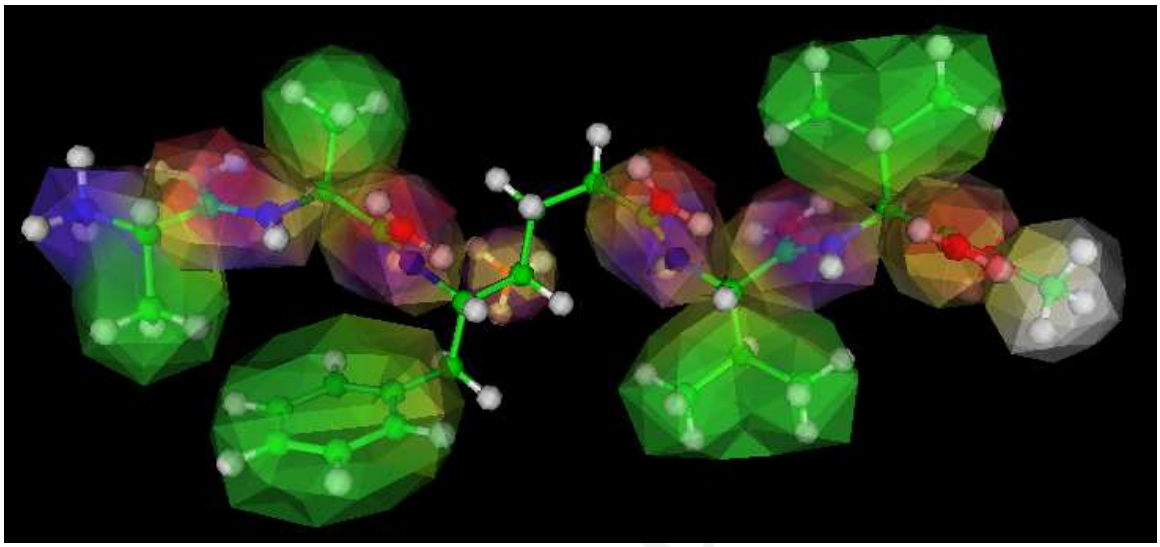


Figure 12: TOC Graphic.