

# eHiTS: An Innovative Approach to the Docking and Scoring Function Problems

Zsolt Zsoldos<sup>1,\*</sup>, Darryl Reid, Aniko Simon, Bashir S. Sadjad and A. Peter Johnson<sup>2</sup>

FINAL

<sup>1</sup>SimBioSys Inc., 135 Queen's Plate Drive, Suite 520, Toronto, ON, M9W 6V1, Canada; <sup>2</sup>Chemistry Department, The University of Leeds, Leeds, LS2 9JT, U.K

**Abstract:** Virtual Ligand Screening (VLS) has become an integral part of the drug design process for many pharmaceutical companies. In protein structure based VLS the aim is to find a ligand that has a high binding affinity to the target receptor whose 3D structure is known. This review will describe the docking tool eHiTS. eHiTS is an exhaustive and systematic docking tool which contains many automated features that simplify the drug design workflow. A description of the unique docking algorithm and novel approach to scoring used within eHiTS is presented. In addition a validation study is presented that demonstrates the accuracy and wide applicability of eHiTS in re-docking bound ligands into their receptors.

## 1. INTRODUCTION

The number of publicly available protein structures in the Research Collaboratory for Structural Biology (RCSB) [1] database has grown to approximately 30,000 structures with thousands being added each year. In addition the number of structures of small molecules available in public databases (such as ZINC [2] or PubChem [3]) or company proprietary databases has reached into the millions. This wealth of available data raises the question of how it can be best used to assist in drug design and discovery

Computational methods, in particular ligand docking programs, have become an essential part of any drug discovery program. The world of docking programs can be broadly divided into two categories, stochastic or random approaches (AutoDock2 [4], DockVision [5], GOLD [6], ProLeads [7], etc.) and systematic or directed approaches (FlexX [8], DOCK[9], FLOG[10], FRED[11], etc.).

Many reviews of docking algorithms and scoring functions have been published in recent years [12-18]. It is commonly reported in these reviews that docking programs are often able to reproduce the correct pose of protein-ligand complexes (alongside many incorrect poses), and that the problem lies in the accurate estimation of the relative binding affinities of ligand poses, i.e. the scoring function.

The ability to reproduce protein-ligand binding poses is therefore a prerequisite for a docking tool to be useful in the drug discovery industry. In this review, results of a validation study using 1629 protein-ligand complexes with the eHiTS docking tool will be presented.

The scoring function problem has been discussed extensively in the literature and is commonly accepted as the main

limiting factor of docking programs. The latest version of eHiTS includes a novel approach to the scoring problem. This new approach will be briefly introduced in this review, however a full discussion of the scoring function will be published elsewhere.

This paper is organized as follows. First an overview of eHiTS docking algorithm is presented followed by an introduction to the new "statistically derived empirical scoring function". eHiTS also contains a pocket detection algorithm that can assist computational chemists find the correct binding site within a receptor, and this pocket detection algorithm will be reviewed. Protonation state is a very important factor in ligand docking, however it is one that is often treated superficially during the docking process. eHiTS automatically evaluates all possible protonation state combinations of the ligand and receptor in a single run, thus eliminating the dependency of the docking program on pre-defined protonation state determinations.

The eHiTS docking algorithm docks all rigid fragments derived from a ligand independently within a receptor site. This means that the results of docking calculations for a particular rigid fragment can be reused in subsequent ligand screens when the fragment is repeated. Section 4 discusses the effect of reusing docking results.

The paper also presents the results of a validation study that highlights the accuracy that has been achieved by the new scoring function.

## OVERVIEW OF DOCKING METHODS

The world of flexible ligand docking, suitable for virtual ligand screening, can be broadly divided into two main categories, systematic methods (incremental construction, conformational search) and stochastic or random methods (tabu search, Monte Carlo, genetic algorithms). Simulation methods such as molecular dynamics and quantum mechanics do

\*Address correspondence to this author at the SimBioSys Inc., 135 Queen's Plate Drive, Suite 520, Toronto, ON, M9W 6V1, Canada; E-mail: zsolt@simbiosys.ca; Web: <http://www.simbiosys.ca>

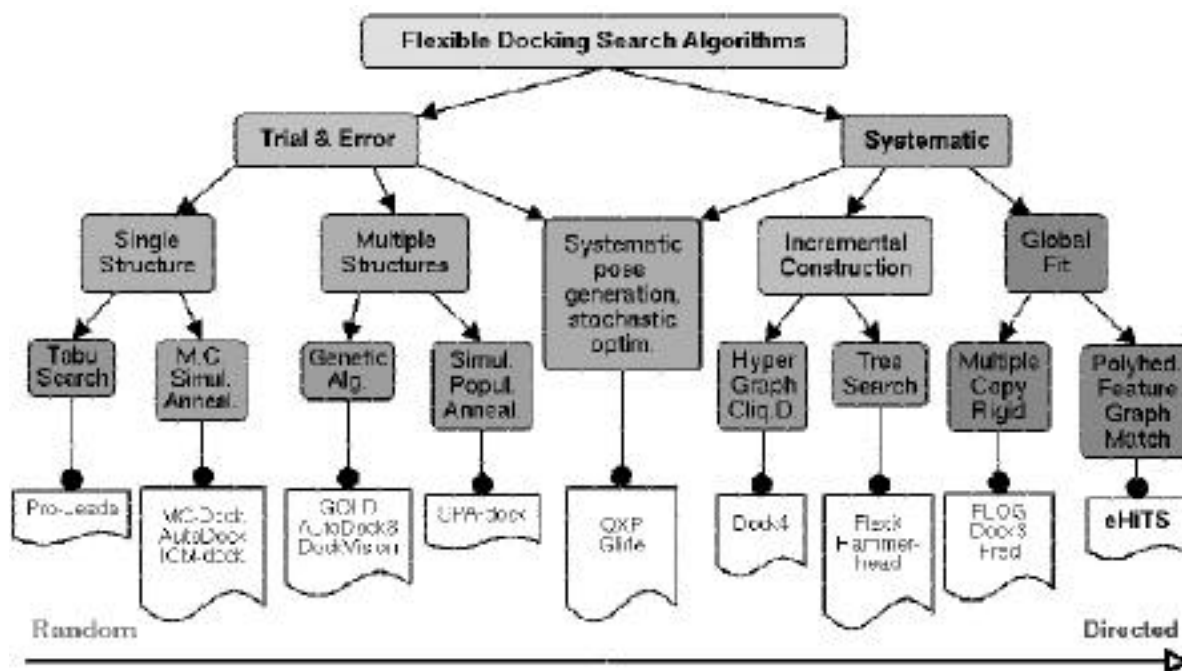


Fig. (1). Classification of several docking algorithms.

exist, but these methods tend to be too computationally demanding to be applied to virtual ligand screening and thus will not be the focus here.

### Random or Stochastic Methods

In Monte Carlo simulations the ligand is placed randomly in the receptor and random changes are applied to the translations and rotations, as well as the torsion angles, of the ligand. After each change the ligand is typically minimized and scored [19]. To improve convergence the simulation usually occurs in several cycles, the first at a high temperature and subsequent cycles at increasingly lower temperatures. This approach is known as Monte Carlo simulated annealing. AutoDock [4] was the first docking program to implement a Monte Carlo simulated annealing algorithm. The docking program ICM [20,21] uses a Monte Carlo Minimization procedure in the internal coordinate space to find the global minimum of the energy function. Each step of the algorithm consists of complete randomization of a single arbitrarily chosen torsion angle and a pseudo-Brownian random translation and rotation of the ligand as a whole.

Tabu search is an algorithm, similar to Monte Carlo, which attempts to improve the sampling of the search space, not by locating a minimum through simulated annealing, but rather by keeping track of previously generated conformations. If a newly random conformation is not lower in energy, it is only kept if it is not similar to conformations on the "tabu" list. PRO\_LEADS [7] uses a Tabu search approach to docking.

Genetic Algorithms are inspired by Darwin's theory of evolution. Ligand conformations are encoded in a chromosome and stochastically varied. Chromosomes are evaluated by a fitness function (scoring function) and are allowed to reproduce. Some chromosomes are allowed to mutate and/or

crossover with others to reproduce and create new individuals. DOCK [22] and GOLD [6] are examples of docking programs that have implemented a genetic algorithm.

### Systematic Algorithms

Systematic algorithms attempt to sample all the possible degrees of freedom of a molecule with some discrete resolution, but face the problem of combinatorial explosion. To deal with this, some docking programs have implemented a stepwise or incremental construction algorithm. In incremental construction ligands are divided into rigid fragments and flexible connection chains. Anchors are typically chosen from the set of rigid fragments and the anchors are docked in the active site. The flexible parts are then docked sequentially with systematic sampling of the torsion angles. DOCK4.0 [23] and FlexX [8] use incremental construction algorithms.

A similar approach is used by the Hammerhead [24] docking program, except that Hammerhead docks all fragments independently and identifies key interacting fragments, or 'heads'. The heads are selected as starting points to rebuild the ligand. The Hammerhead approach differs in that it performs an energy minimization to optimize torsions after each fragment is added. However fragments are still added in an incremental fashion.

The docking program eHiTS also docks all fragments independently within the receptor site. However, rather than selecting 'heads' upon which to incrementally reconstruct the input ligand, eHiTS uses a hypergraph matching algorithm to enumerate all compatible 'pose sets'. A pose set is a set of fragment poses (one pose for each fragment) that are capable, in terms of distances between fragments, of reforming the input ligand. This approach makes the search exhaustive, i.e., it finds all possible solutions. In addition to

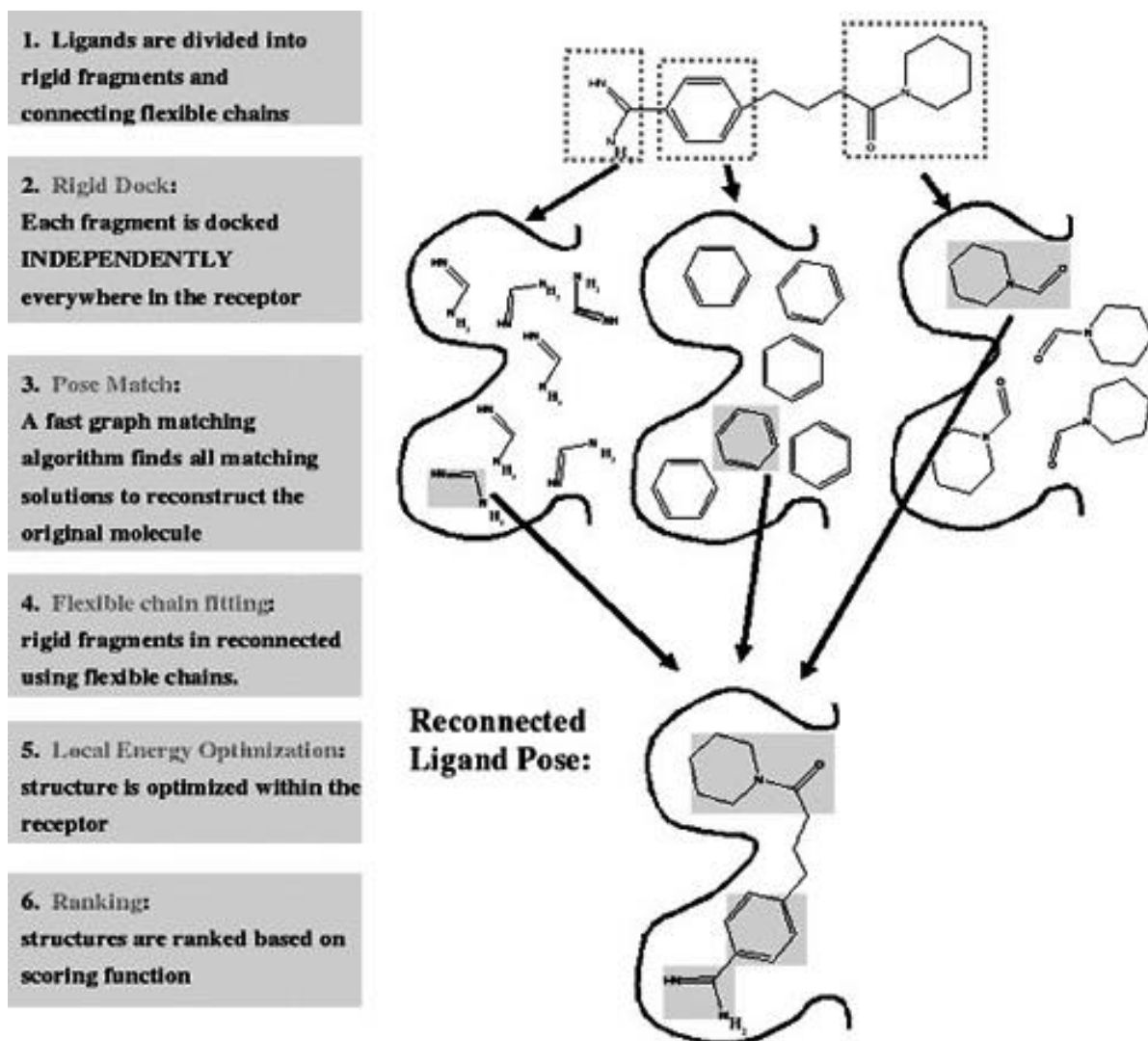


Fig. (2). Overview of the eHiTS docking algorithm.

eliminating seed bias, decisions on which poses to retain for further processing and optimization, are made based on a global score of the full ligand, and not on partial structures as in incremental construction algorithms. Also since in the eHiTS approach, the docked position of all fragments of a single pose are known prior to reconnecting the rigid fragments, there is no requirement for discrete sampling of torsion angles.

## 2. THE eHiTS METHOD

eHiTS takes a unique approach to the docking problem, both in its innovative docking algorithm and novel approach to the scoring function problem. The eHiTS approach involves breaking ligands into rigid fragments and connecting flexible chains and systematically docking each rigid fragment to every possible place in the cavity independently.

An exhaustive matching of compatible rigid fragment pose sets is performed by a rapid graph clique detection algorithm, resulting in a few hundred thousand (small pocket, few rigid fragments) to several million (large pocket, many

small rigid fragments) acceptable combinations of poses. 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 combinations are the best. This means that even fragments that had poor interaction scores with the receptor, may form part of a pose set that scores very well, and the full pose set would be accepted. This is frequently the case with linker fragments.

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.

### 2.1. Geometric Shape and Chemical Feature Graph

The fragmentation of the ligand is focused on separating rigid fragments from the flexible linkers. eHiTS identifies rotatable bonds within the input ligand, and these bond are removed leaving behind a set of rigid fragments. Whenever a

bond is broken during the fragmentation process, both atoms of the bond are duplicated, that is to say they are kept in both the rigid fragment and the flexible chain. The distance between these “*join atoms*” is used to determine the compatibility of rigid fragment poses in the pose-match phase of the algorithm.

All ring systems are considered rigid and their conformations are preserved, therefore it is desirable to use multiple ring conformers for complete conformational sampling. Acyclic fragments with double or normalized (resonance) bonds and all sp<sup>2</sup> hybridized atoms are considered rigid, e.g. including the amide functional group.

Both the receptor cavity and the candidate ligands are described by a Geometric Shape and Chemical Feature graph (GSCF). The nodes of the GSCF graph represent a rigid shape by a simplified geometric hull. This is derived from regular polyhedra and then distorted to shrink-wrap the actual molecular fragment or cavity region. Chemical properties are associated with each vertex of the polyhedron.

## 2.2. Rigid Fragment Docking

Each rigid fragment is placed in each cavity polyhedron during the rigid docking phase of the eHiTS algorithm. For each cavity – rigid fragment pair, all orientations of the polyhedron are explored. The polyhedron representation allows for very rapid enumeration of all fitting poses. Each vertex on the rigid fragment and cavity polyhedra has a surface point type associated with it (i.e. its chemical properties). A scoring matrix is defined for each identified interaction pair (see the section on scoring for more information). The score of a rigid docking pose is computed by summing all the applicable scores of any interacting surface points between cavity and ligand.

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).

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 contained fragments that either do not make any interaction with the protein, or even make interactions that are clearly repulsive. 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.

A key advantage of this approach is that since *all* acceptable poses for each rigid fragment are computed regardless of other fragments in the ligand; the information about a particular rigid fragment can be reused in a subsequent docking when that particular rigid fragment is present again. This situation occurs very frequently during a virtual screening study when many thousands or even millions of drug-like molecules are docked to a given receptor. 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. More information about the use of the database in eHiTS can be found in section 4.

## 2.3. Pose Matching

There are several thousand alternative poses generated and scored at the rigid docking step for each rigid fragment. Several hundred poses for each rigid fragment are passed to the next stage of the algorithm, based on a clustering algorithm. 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. This is solved by clique detection on the following graph. Each rigid fragment is represented by a set of graph nodes; one corresponds to every accepted rigid fragment pose. There are edges between those node pairs where all the following conditions hold true:

- the nodes correspond to poses of different ligand fragments
- there is no steric clash between the two poses
- if the fragments are connected by a chain, 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 graph should consist of as many nodes as the number of rigid fragments in the ligand. 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.

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 computational 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.

## 2.4. Flexible Chain Fitting

The next stage in the eHiTS docking algorithm is the fitting of flexible chains that connect the rigid fragments. 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 most suitable low energy chain configuration is selected from a lookup table based on the relative 3D positions of the desired end points.

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 is able to produce any dihedral necessary to reach the end points and resolve clashes – if necessary, it will even allow the highest local energy

eclipsed conformation. 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 endpoints 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. Therefore the dihedral angle sampling of eHiTS is practically equivalent to continuous sampling.

### 3. SCORING

Estimating binding affinities of ligands within a receptor is a challenging task that is crucial to virtual ligand screening. Free-energy minimization techniques have been developed for quantitative modeling of protein-ligand interactions and prediction of binding affinities [25,26]. However these approaches are far too time consuming to be applicable to virtual ligands screening. Therefore, more computationally cost effective approaches have been developed. Scoring functions in docking programs make assumptions and simplifications in the effort to reach a balance between computational time and accuracy of the results. Essentially there are three classes of scoring functions used in docking programs, force-field based, empirical and knowledge based. Fig. 3 shows a classification of some of the key scoring functions used in docking programs.

Knowledge base scoring functions use statistics collected from experimentally determined protein-ligand complexes to extract rules on preferred and non-preferred atomic interactions. They are designed to reproduce binding poses rather than binding energies. Rules are interpreted as pair-potentials that are subsequently used to score ligand binding poses. Common examples of knowledge based scoring functions include PMF[27-29], DrugScore [30], and SMOG [31].

Empirical scoring functions consist of the sum of a set of functions parameterized to fit experimental data, such as binding energies or conformations. The idea is that binding energies can be approximated by a sum of individual uncorrelated terms. The weights of these terms are assigned by regression methods that are used to fit the experimentally determined values found in a training set of protein-ligand complexes. The interaction terms typically have some physical meaning, such as Van der Waals, electrostatics interactions and hydrogen bonds. ChemScore [32], LUDI [33,34], F-Score [35], SCORE [36,37], X-Score[38] and Fresno[39] are all examples of empirical scoring functions.

Force-field based scoring functions are similar to empirical scoring functions, in that they attempt to predict binding energies of ligands by adding individual contributions from different types of interaction. However, force-field based scoring functions use interaction terms derived from physical chemical phenomena as opposed to experimental affinities. Some examples of force-field based scoring functions include D-Score [40], G-score [39], GOLD [41], AutoDock [42] and DOCK [43].

Many reviews have been published on the use of various scoring functions [44-48] in docking programs. The general consensus is that there is currently no universal scoring function. However, in work by Wang *et al.*, [49] a comparison was performed on 11 scoring functions on their score-rank ability on the same set of configurationally diverse ligand poses (generated by AutoDock) for 100 receptor-ligand complexes. The authors showed that on this test set 6 of the 11 scoring functions were better than the native AutoDock scoring function. They concluded in general, empirical scoring functions performed better on this set, while force-field based scoring functions were the weakest. In addition they demonstrated that the performance of different scoring functions is influenced by the nature of the protein-ligand

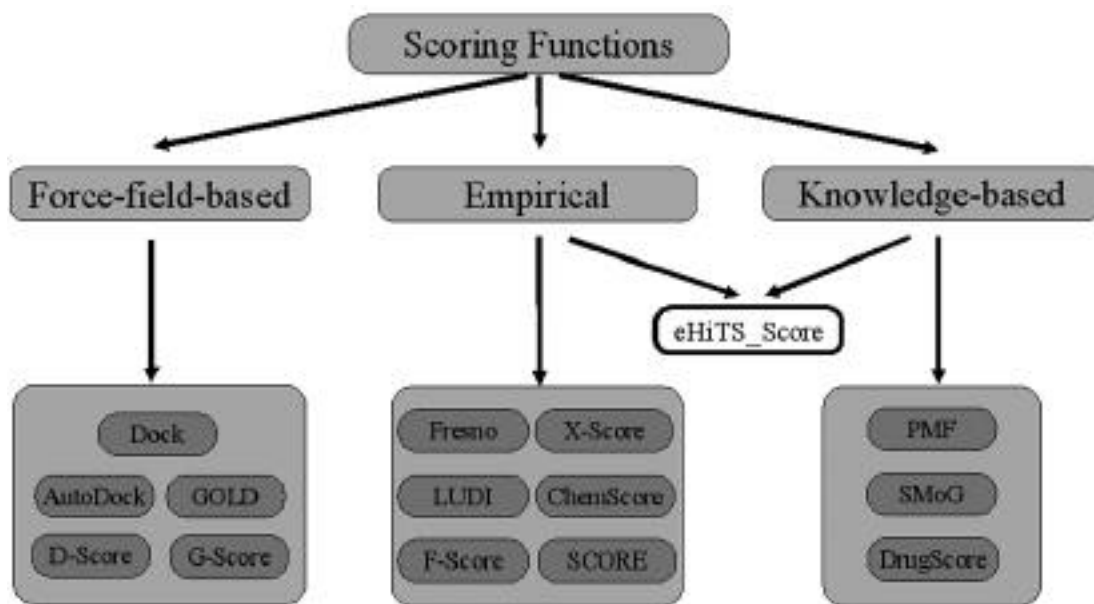


Fig. (3). Classification of several scoring functions.

interactions. Miteva *et al.*, [50] also concluded that scoring functions should be developed to direct them toward specific binding sites or that users tune the parameters to reflect the receptor of interest.

The docking program eHiTS takes a unique approach to the scoring function problem. eHiTS uses eHiTS\_Score, a statistically derived empirical scoring function, with many novel features, including the fact that it takes into account the temperature factors of crystal structures. In addition to a generic, default weight set, eHiTS\_Score uses automated receptor family clustering which is used to create sets of family-specific scoring function weights which better represents those receptors. While a full description of the scoring function will be published elsewhere, a general description will be presented here.

In standard PDB files, there is a temperature factor (B) associated with the position of each atom. That factor correlates to the probability distribution of that atom around the stated coordinates. The atomic coordinates stated in a PDB file is an average of all "observed" poses of the protein in the crystal used for the structure generation. Each atom has different positions in each copy of the protein within the crystal. Some have a very precisely confined position, while others are more loosely defined. Any treatment which ignores this and treats all atom coordinates as equally significant fails to take advantage of all the information provided. For statistics collection on, for example, H-bond geometries, it is very important to recognize which arrangements are well defined with low temperature factors and which geometries are mere averages of wide variations. Some geometries occur with high frequency, but if they always occur with high temperature factors, then it does not mean that the specific geometry is really preferred.

The scoring in eHiTS is based on interaction of surface points. The rigid fragments of the ligand are "shrink-wrapped" with a surface, and the vertices are assigned chemical properties. Similarly, the receptor binding site is filled with shapes, centered on a 0.5Å grid, that also have chemical properties associated with the vertices. The interaction between ligand surface points and receptor surface points determine the score given to that rigid fragment, and ultimately the score for the entire ligand.

The relationship between surface points cannot be fully described by just the distance between the atoms associated to those surface points. The angle between the surface points and the line between the atom, as well as the torsions between the two surface points are also important to the interaction. To clarify this, consider a hydrogen bond from an -OH in a ligand and a nitrogen on a receptor, see Fig. 4. In

(Fig. 4) (a) the hydrogen makes a good interaction with the lone pair on the nitrogen. However, if the torsion was changed by 180°, there is no interaction, as depicted in (Fig. 4) (b). In this situation the atom separation is the same, the angles are the same, but the interaction is completely different. Therefore all four variables are required to describe the interaction.

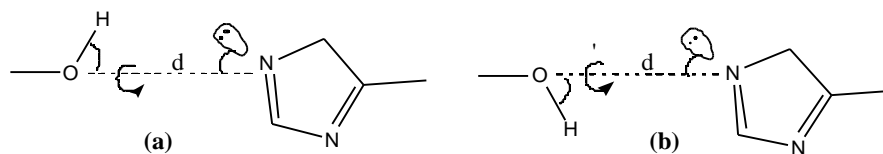
Based on the assumed Gaussian distribution of atom positions and the mean coordinates given in the PDB files, the probability of any given interaction geometry can be expressed with a volumetric integral that depends on the 4 variables (distance, 2 angles and a dihedral). Therefore every individual atom-atom interaction observed produces a continuous 4D probability field.

Interaction statistics were gathered using a set of 1420 high resolution (2.5Å or better) protein-ligand complex PDB files for which the receptors were saved in PDB format and the ligands were saved in MOL2 format. The interaction statistics were collected for all pairs of atoms within 5.6Å of each other, except for those that are connected by a covalent path of less than 4 bonds (i.e. direct, secondary and tertiary neighbor atoms within the same molecule are not considered as interacting pairs). There were about ten thousand individual interactions per complex on average, giving a total of about 10 million interacting atom pairs as the input to our statistics collection.

The surface points are classified into 23 distinct types (analogous to atom types in other scoring functions or force fields) that are listed in Table 1. The interaction statistics are collected separately for every different surface point type pair, i.e. data is collected in a matrix of 23x23 (only half of the matrix is used due to the symmetrical nature of the matrix, i.e. 23x24/2 cells). Each cell of this matrix contains a 4D probability function summarized from the input statistics using the volumetric integrals.

Four variable third degree polynomials are fitted using a partial least square fit to the collected data (in each cell separately) to enable fast and continuous scoring function evaluation. The polynomials have the same generic form for each cell, but individually fitted 20 parameters. So the final usage of the scoring function is not a statistical lookup, but the calculation of a polynomial formula much like a typical empirical function. However, the 20 parameters for each of the 23x24/2 different interaction type pair are derived from statistical data, hence this scoring function should be categorized as a "statistically derived empirical scoring function".

To use this scoring function in the docking problem, we must be able to convert the interaction probability into energy. The random probability of interaction was determined



**Fig. (4).** The relationship between ligand atoms and receptor atoms are determined by four variables,  $d$ ,  $\alpha$ ,  $\beta$ , and  $\gamma$ , (a) shows a good hydrogen bond interaction however in (b) we see the same distance ( $d$ ) and angles ( $\alpha$ ,  $\beta$ ) but by changing the torsion ( $\gamma$ ) there is no longer a good interaction formed.

**Table 1. The 23 Surface Point Types as Defined in eHiTS**

Surface Point Type	Definition
METAL	positively charged metal ion point
CHARGED_HPLUS PRIMARY_AMINE_HLP	positively charged hydrogen, e.g. Arginine primary amine hydrogen/lone-pair, e.g. $-\text{NH}_3^+$ or $-\text{NH}_2$
HDONOR	strong (primary) hydrogen bond donor H (polar-atom-H)
WEAK_HDONOR	weak (secondary) hydrogen bond donor H (polarized C-H)
CHARGED_LONEPAIR	lone pair of negatively charged group, e.g. $\text{PO}_3^-$
ACID_LONEPAIR	lone pair of an acid group, e.g. carboxylate
LONEPAIR	strong (primary) hydrogen bond acceptor lone pair
WEAK_LONEPAIR	weak (secondary) hydrogen bond acceptor lone pair
AMBIVALENT_HLP	donor H OR acceptor Lp depending on protonation state
ROTATABLE_H	rotatable-hydroxy donor H
ROTATABLE_LP	rotatable-hydroxy acceptor Lp
HYDROPHOB	H on aliphatic (chain) hydrophobic carbon
H_AROM_EDGE	H on hydrophobic carbon in aromatic ring (non-polarized)
WS_LIPO	H on weak secondary hydrophobic atom (e.g. carbon next to polar)
NEUTRAL	H/Lp on neutral atom (no recognized activity)
PI_AROMATIC	electron of an aromatic ring
PI_RESON_POLAR	electron on polar atom (N/O) in resonance chain, e.g. amide
PI_RESON_CARBON	electron on carbon atom in resonance chain, e.g. amide
PI_SP2_POLAR	electron on $\text{sp}^2$ polar atom (N/O) (non-resonating, non-arom)
PI_SP2_CARBON	electron on $\text{sp}^2$ carbon atom (non-resonating, non-arom)
HALOGEN	lone electron pair of a halogen atom (F,Cl,I)
SULFUR	lone electron pair of a sulfur atom

and an energy scaling factor was assigned to each cell of the interaction matrix based on the Boltzmann equation. These scaling factors are collected and set once and is the basis for the family-based scoring functions described below.

In addition to the derived scoring functions and determined energy scaling factors, the final scoring function also includes terms for

- steric clash (quadratic penalty function),
- depth value within binding pocket,
- receptor surface coverage (exposed/buried hydrophobic area),
- family-coverage,
- conformational strain energy of the ligand,
- intra-molecular interactions within the ligand,
- entropy loss due to frozen rotatable bonds

### 3.1 Training the Scoring Function

The terms of the scoring function are combined using adjustable weights. The  $23 \times 24/2$  functions created in the statistics phase are mapped to 13 weight factors, to allow for easier modification of the weights. Rather than having a single weight set for all kinds of proteins, better scoring accuracy can be achieved by using a different weight set for each protein family. The categorization of different proteins into different families and the tuning of the weight set for each family are completely automated processes. The family training starts with collection of receptor site descriptor information followed by the clustering of the receptors into families based on the collected information. The descriptor information consists of

- A collection of surface point types that are on the interaction surface of the receptor cavity (as identified by the pocket detection algorithm)

- b) Statistics on the ligand surface point types that interact with each of the receptor surface points (frequency and geometry)
- c) The residue names and numbers of the heavy atoms corresponding to the surface points
- d) distance matrix of the centers of the residues that contain the above atoms

The clustering is based on having the same residues appear with compatible distance matrix in different protein structures. The number of required matches is a user-adjustable parameter with the default value of 5, meaning that 5 residues have to match with all their respective distances from each other ( $5 \times 4 / 2 = 10$  residue-pair distances) within a tolerance of 3Å.

If a specific residue matches more than 1 existing family, then the best match is selected (highest number of residues matching within tolerance, if that is still the same then the more precise distance match counts).

The statistical information (b) is summed up for all the surface points in the family and used in the family-coverage scoring component.

It is important to note, that the family clustering is not based on biological activity information or naming or references in the PDB files, but purely on the residue types that form the active site and the shape of the active site (described by the distance matrix of all residues forming the active site). Thus our algorithm may discover binding site similarities automatically without the need for manual classification.

For the training/tuning of the family specific weight sets, 884 receptor-ligand complexes were selected from the 1420 complexes used during the statistics collection phase based on the Lipinski rules for drug like ligands [63]. The 884 complexes were clustered, automatically, and generated 71 clusters or families, ranging from 2 to 151 complexes per family. Families with less than 5 complexes were augmented

with other complexes from the PDB, which also contained 'drug-like' ligands.

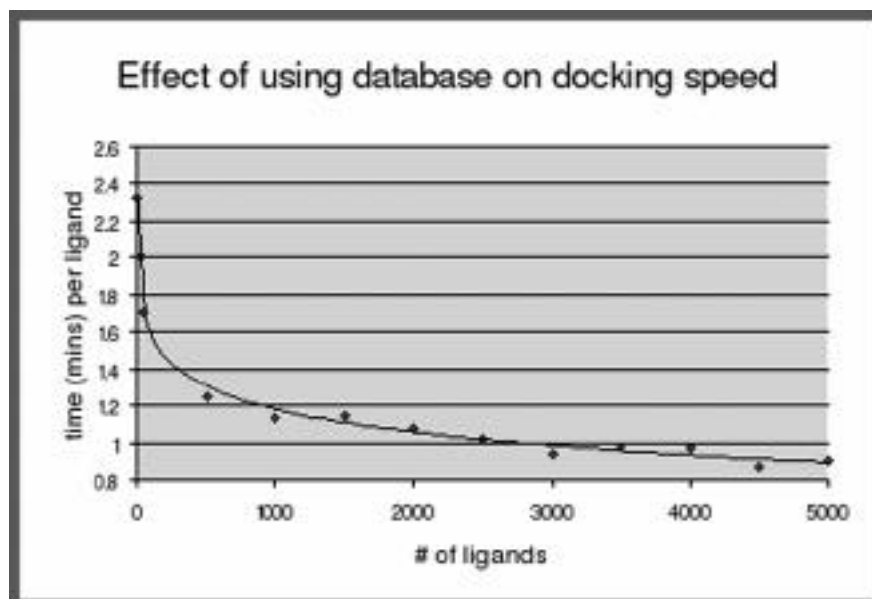
The augmentation resulted in a final training set of 1315 complexes used during the training phase of the scoring function generation. 346 of these were not clustered into specific families (i.e. they were singletons) and were used to train the default parameters of the scoring function. Thus, in its final form the eHiTS Score scoring function contains 72 weight sets, 71 family-based and 1 default weight set to be used for unrecognized receptors. Results, shown in detail in section 6, show that family based training improves both docking accuracy and screening (enrichment).

#### 4. RE-USING DOCKING RESULTS IN eHiTS VIA THE DATABASE

Due to the fact that eHiTS docks all rigid fragments in a ligand independently, everywhere within a receptor site, it is possible to re-use the docking information for subsequent ligands if the fragment is repeated. This is an extremely useful and time saving feature of eHiTS. There are many rigid fragments in "drug-like" molecules that are repeated many times in a database of ligands. By storing the docking information for that fragment in a database, eHiTS is able to simply read that information rather than having to recalculate it.

The docking of rigid fragments depend on the receptor binding site, therefore the results cannot be reused for different receptors, or if the binding site specification has changed. In this case a new database would have to be created,

To test the effect of the database on a virtual ligand screen, 5000 ligands were screened against an estrogen receptor (pdb code 1ERR). The 5000 ligands were randomly selected from a set of drug-like ligands obtained from the ZINC [2] database of ligands. Using the standard mode of eHiTS, 50 ligands were screened without using the database, to judge the baseline time per ligand. For this case, eHiTS averaged 2:20 minutes per ligand on a P4 3.06GHz cpu. The



**Fig. (5).** The time per ligand drops from 2.4 mins/ligand to under a minute per ligand as the database is filled with rigid fragments.

plot below, (Fig. 5) shows the speed-up due to use of the database during the eHiTS screen.

It can be seen that as the database fills up, i.e. more ligands are run, the speed to dock each ligand decreases, and at 3000 ligands the speed is just under 1 minute per ligand, compared to 2:20 minutes without the database. The speed-up tends to level off around 3000 to 4000 ligands, meaning that most of the repeating rigid fragments have already been placed in the database. Therefore eHiTS has a limit on the number of unique rigid fragments that is stored in the database (default value is 10,000). By preventing the database from getting too big the speed-up performance is maintained.

## 5. POCKET DETECTION

The identification of small molecule binding sites in proteins is essential to any VLS application. For many VLS experiments, the location of the binding site is known, either from a co-crystallized protein-ligand complex or some other means. However, in some instances this information is not available and thus other approaches must be used to select the correct binding site for docking. Binding site identification methods have recently been reviewed by Sotriffer and Klebe [51], as well as Campbell *et al.*, [52].

There are two main approaches to binding site identification, docking or probing techniques and geometric algorithms. Methods that use docking or probing techniques scan the surface of the protein for areas of high complementarity with respect to certain molecular fragments or entire ligands. The approach by Ruppert *et al.*, [53], for example, coats the protein with molecular fragments or probes. The position of each fragment is scored to give an estimate of the binding affinity. The binding site is then detected by looking for regions of high affinity and high density. Other similar methods include the SuperStar algorithm [54,55], the vdW-FFT method of Bliznyuk and Gready [56,57] and the more recent Q-SiteFinder [58].

Geometric approaches examine the surface of protein structures for clefts or pockets. LIGSITE [59], based on the earlier POCKET program [60] places the protein in a regular Cartesian grid and scanning along the x, y and z axes and the cubic diagonals for areas that are enclosed on both ends by the protein. SURFNET [61] uses spheres placed between protein atoms to fill up the clefts in the protein, this creates areas of interpenetrating spheres which correspond to the protein's cavities or pockets. APROPOS [62] and CAST [63] are based on the alpha shape algorithm, they compare surfaces of a protein generated at different levels of resolution to identify pockets.

The automatic pocket detection algorithm of eHiTS is based on pocket-depth values associated with 3D grid cells around the receptor structure. The grid is generated in a 3D box that encloses the whole receptor (alternatively, it can be limited by a user-defined clip-box). The resolution of the grid is 0.5Å in each dimension.

A qualitative numerical definition of the depth value at any surface point is computed using the following multi-phase flood fill algorithm:

The 3D grid is initialized to indicate which positions are available to ligand atoms based on the Connolly surface of the receptor, i.e. any grid cell within the Connolly surface is

considered occupied (not available for ligand atoms). The grid cells that fall within 0.25Å from the analytical Connolly surface are collected and referred to as "surface points".

A flood fill is started from all surface points outward from the receptor to determine the smallest distance of any grid point from the receptor surface. This flood stops at the bounding box of the grid. The values at the box are approximated distances from the convex hull of the receptor.

A second, negative flood is started from the bounding box (with initial value equal to the value assigned in step number 2 above) marking grid cells with decreasing distance values until the surface points are reached. The values would reach zero at the approximated convex hull of the receptor. Then the values descend to negative range inside the concave parts, i.e. deeper in the pocket. These depth values are re-membered for each grid cell and used by the scoring function of eHiTS.

A third flood is started from the deepest negative value, which indicates the depth of the deepest pocket. This flood determines what volume belongs to the same pocket as the deepest point. It is not allowed to flood to positive depth values. With this flood, the extent of the pocket is identified and cut off from any other pockets that are not linked with the deepest. This step can be repeated for any remaining (not previously covered) local minimum depth value, so that all independent pockets are identified.

This algorithm has been tested on over 300 PDB complexes. Typically it identifies 2-4 independent pockets on a protein receptor. In most of the cases (about 85%) the deepest pocket is the largest in volume and the co-crystallized ligand is located in that pocket, i.e. it is the correct binding site. For the remaining cases the pocket with the largest volume is not the same as the deepest, but one of those two is the correct binding site. We have implemented a simple heuristic selection algorithm which considers the depth values, volumes and their respective ratio to decide which one is probably the binding site. With the heuristic rules, the correct pocket is identified in 99% of the test cases.

The importance of the depth values computed by the algorithm for any point inside the volume will be highlighted in the scoring section.

The algorithm is very fast, and does not rely on chemical perception. This feature allows the eHiTS software to be used at the leading edge of drug discovery science, i.e. to dock potential drug candidates against new therapeutic targets whose function and active site is not yet known. With the completion of the human genome project and the advancements in protein folding and homology modeling, the number of determined 3D protein structures has grown rapidly. However, the location of the active site is unknown for many of these structures, because co-crystallized complex structures are not available. Automatic pocket detection makes it possible to find good ligands for these new targets. Clearly these ligands are tailored to the cavity of the protein

lacking any ligand, and it is recognized that the presence of ligands may change the shape and size of the cavity by induced fit.

## 6. AUTOMATIC PROTONATION STATE EVALUATION

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. A better solution, with a more appropriate score, can only be found if the program is run with different 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 (Fig. 4) has 200 possible protonation states. Table 2 shows the 5 possible protonation states for each of A and B, 2 for B and 3 for C, combined this leads to  $5 \times 5 \times 2 \times 4 = 200$  different possible protonation states. Although, two pairs of states for A&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 fragments giving a total of  $3 \times 3 \times 2 \times 4 = 72$  instead of 200. Most docking programs would need to dock all 200 (or at least 72) 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. eHiTS systematically evaluates all possible protonation states for the receptor and ligands, automatically for every receptor-ligand pair. It does this through the use of ambig-

uous properties flags for positions that could be either protonated or deprotonated (i.e. have a lone pair). Then during the docking algorithm each state is evaluated and scored, selecting the best protonation state for each individual interaction without the combinatorial effect. The result is a docking program that evaluates all possible protonation states for the receptor and ligand in a single run.

Fig. (5) shows an example of how the chemical properties (i.e. surface point types) are applied to the ring of a histidine residue. Both nitrogen atoms have the same surface point type towards the edge of the ring (Ambivalent H/LP) which indicates that it may be a donor H OR acceptor Lp depending on protonation state. This simultaneously handles all three protonation states of histidine. See the scoring section for more information on how the surface point types are used in the scoring.

## 7. VALIDATION STUDY

In selecting the validation set to illustrate the accuracy and wide-applicability of eHiTS, it was decided that a cross section representation of different receptor families would be chosen. The validation set consists of two main sets, a set of 884 complexes used as part of the training set of the eHiTS scoring function and a set of 742 complexes downloaded from the PDB that are not part of the training set. To our knowledge this validation set of 1626 complexes is the largest validation set published to date.

The 884 complexes from the training set (here on referred to as the 'training set'), consists of all 'drug-like' ligands, as defined in the Lipinski [64] rules. Each complex in the training set was split into a receptor molecule, saved in PDB format, and a ligand molecule, saved in MOL2 format. Metal ions, if present, were kept with the receptor, as were any inorganic and organic cofactors. All water was removed. Hydrogens were not added, as eHiTS handles protonation state automatically during a docking run. As mentioned in the scoring section of this review, eHiTS Score uses a family-based approach to training the scoring function. The 884 drug-like training set was clustered using the eHiTS receptor clustering algorithm. This resulted in 71 receptor families, from 538 complexes. The remaining 346 (884-538) drug-like complexes from the training set were used to train the default (generic) scoring function of eHiTS.

In addition to the training set, a new set of 742 com-

**Table 2. Possible Protonation states for the Functional Groups Identified in the Molecule in Figure 4**

A & D	
B	
C	

**Table 3. Size and Composition of the Training and Test Sets Used in the eHiTS Validation Experiment. <sup>a</sup> The Number of Families Represented in the Set**

	Training set	Test set
Complexes in Families	535 (71) <sup>a</sup>	161 (50) <sup>a</sup>
Complexes not in Families	349	581
Total	884	742

plexes were chosen in the following manner. First 2000 randomly selected complexes with resolution less than 3Å were downloaded from the PDB [1]. These complexes were run through the split utility of eHiTS, which splits the ligand from the receptor and removes the waters. 315 complexes gave problems in this process. An additional 303 complexes had ambiguous ligands (split identified multiple possible ligands) and were therefore removed from the test set. The 1382 remaining complexes were screened for drug-like ligands and resulted in 742 complexes that were used in the validation study (referred to as the test set). The ligands in the test set were not converted into MOL2 format, as was done in the training set. This was done to show that eHiTS is capable of automated processing of PDB files directly with no manual intervention.

No optimization was performed on the input ligand or receptor. For all results shown the standard (default) parameter set was used. No manual preprocessing was performed on any of the PDB complexes. The protonation state, cofactors, counter-ions, solvent molecules, 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 and the accuracy measured by calculating the symmetry-corrected 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. Results are

reported for the “Top ranked pose”, which is the RMSD of the best scoring pose, as ranked by eHiTS Score, and for the “Best found pose” which is the pose with the lowest RMSD from the x-ray pose within the top 32 poses reported.

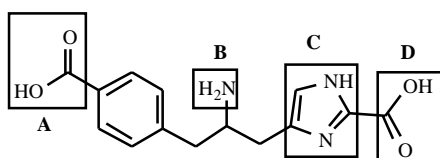
Table 4 shows the average RMSD for the top ranked and closest pose found compared to the x-ray pose of the ligand for both the training and test sets. Overall the training set performs better, as is expected, giving an average top-ranked RMSD of 1.95Å compared to 2.58Å for the test set. However when the family trained results are examined, the difference is only 0.2Å, dropping from 1.69Å to just 1.89Å. If we look at the percentage of structures found under 2.0Å, see (Fig. 9), the test set has 71% of the top-ranked structures docked, compared to 75% for the training set. So while it does appear that the results are slightly better for the training set, the results are close enough to alleviate any concern of over training.

It is also important to note that for both the training set and the test set over 90% of the time there was a pose found under 2.5Å, in the top 32 poses,. This result is true for both the family recognized complexes and the unrecognized complexes. This speaks to the exhaustiveness of the eHiTS docking algorithm.

Comparing the family recognized test set to the unrecognized set, we can see that the family recognized complexes dock with 71% of the top-ranked poses with RMSD under 2Å, while the unrecognized set had only 45% of the top-

**Table 4. Summary of the Average Results for the Training and Test Sets. All RMSD Values are in Angstroms. The Top-Ranked Pose Refers to the Pose Scored the Best by eHiTS\_Score, While the Best Pose Refers to the Pose in the Top 32 Poses that was Closest to the X-ray Structure**

	Test set	Training set
	Top ranked pose	
Average RMS ALL	2.58	1.95
Average RMS Family tuned	1.89	1.69
Average RMS No family	2.77	2.35
	Best pose found	
Average RMS ALL	1.18	0.86
Average RMS Family tuned	0.97	0.75
Average RMS No family	1.23	1.03



**Fig. (6).** Sample ligand with 200 different possible protonation states. Functional groups A, B, C and D have various different forms important to docking.

ranked poses under 2Å. This shows that the family based scoring functions produce more accurate docking results. The 161 family recognized complexes were run using the default, globally trained scoring function, and the results were similar to the unrecognized results (data not shown).

## 8. VIRTUAL LIGAND SCREENING STUDY

To illustrate the effect of family based trained scoring function on virtual ligand screening, it was decided to use some publicly available data from Cummings *et al.*, [65]. Cummings *et al.*, compared four docking tools, DOCK, DOCKVISION, GOLD and Glide on their screening ability over five target proteins, human immunodeficiency virus protease (HIV-Pr), protein tyrosine phosphate 1b (PTP1b), thrombin, urokinase plasminogen activator (uPA), and the human homologue of the mouse double minute 2 oncoprotein (HDM2). However for uPA and HDM2 unpublished structures were used, therefore this data was not available for our study.

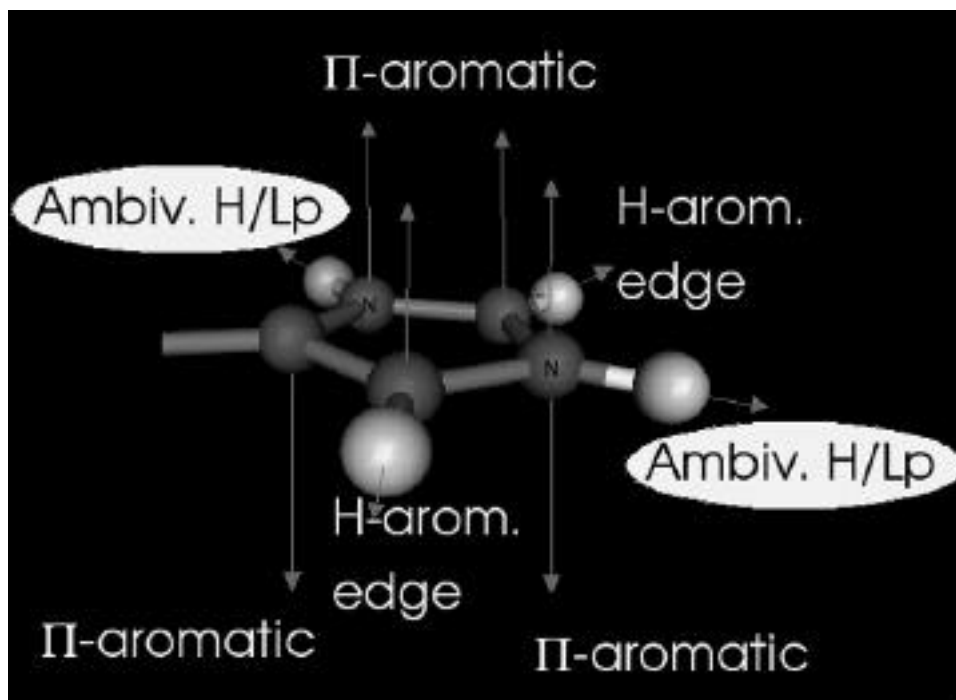
The PDB identification codes for HIV-Pr, PTP1b and thrombin used in this study are 1HVP[66], 1C84[67] and 1QBV [68], respectively. 5 actives for HIV-Pr, and 10 for

each of PTP1b and thrombin were used in the ligand screens presented, along with 1000 randomly selected MDL Drug Data Report (MDDR) [69] compounds as presumed inactive molecules. Full details of the actives and presumed inactive molecules can be found in Cummings' paper [14].

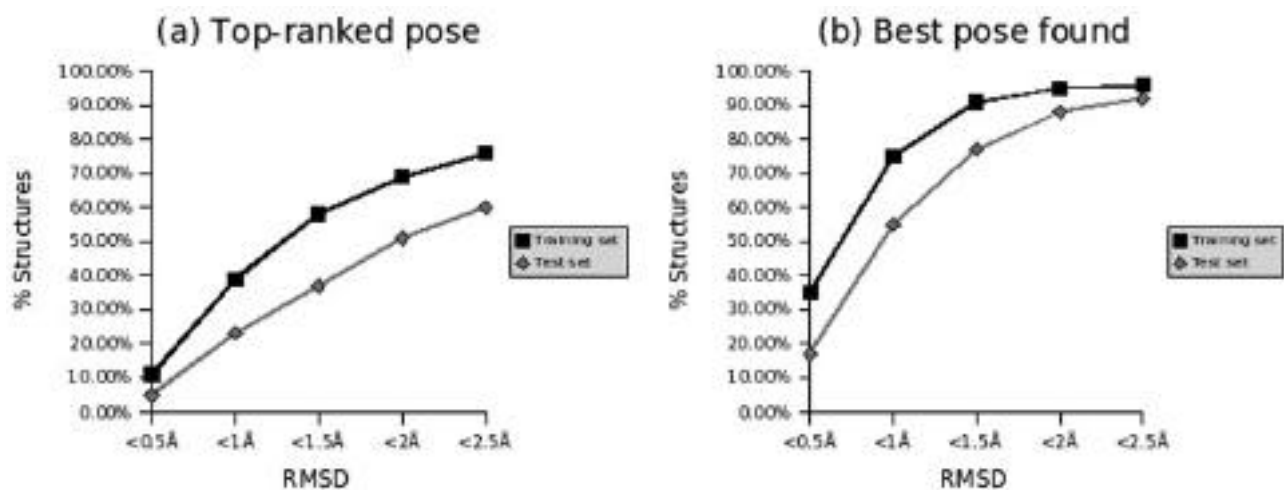
To illustrate the effect of family-based training on enrichment results, two screening runs were completed for each protein target. The first used the unbiased, globally trained eHiTS Score scoring function (marked eHiTS unbiased in the plots in Fig. (11)). The second used the family-based scoring function for each of the identified receptors. The family trained results are shown as solid lines and labeled eHiTS Family in Fig. (11). The results show clearly that the family trained results achieved greater enrichment results than the default, globally tuned, scoring function.

It is clear that the family trained eHiTS scoring function performs much better on the three receptors screens than the default, globally trained eHiTS scoring function. This agrees with the results of the validation run. Family trained scoring functions better represent the characteristics of the receptor binding site than the globally trained weight set.

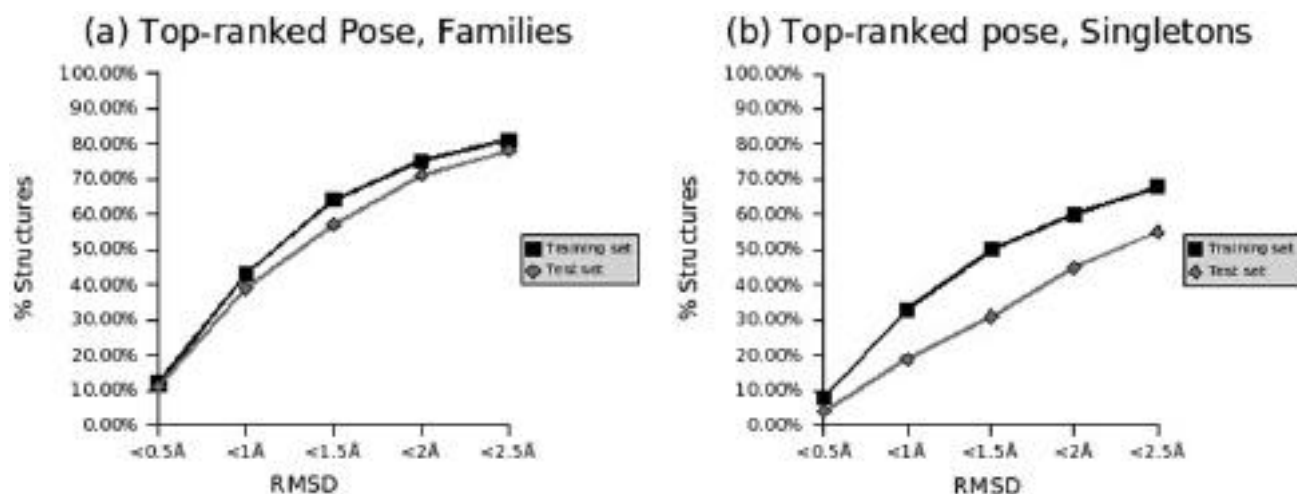
The results for HIV-Pr illustrates the dramatic effect that family based training can have on screening results, however the effect on the other two receptors were slightly muted. For thrombin, this may be due to the fact that the "family cluster" to which thrombin belongs also contains Factor Xa, Trypsin, and Urokinase receptors, thus making the training of this particular weight-set less specific to thrombin. The same cannot be said for the PTP family. This family had 12 example complexes in the training set. Further investigation will be required to determine the behavior of this family.



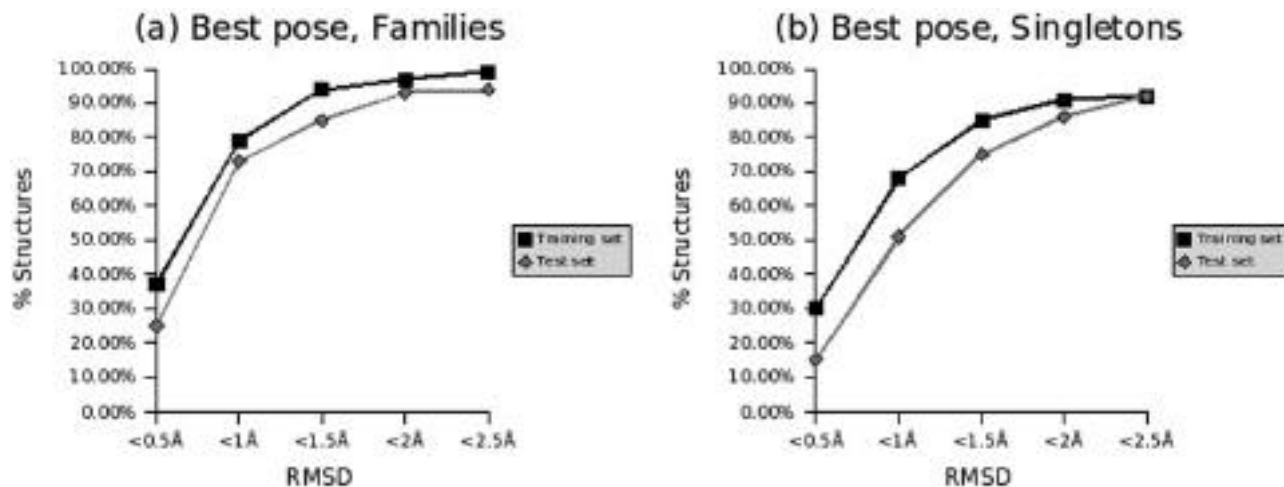
**Fig. (7).** Illustration of how surface point types are applied to the ring portion of a histidine residue. Note the edge points at both nitrogen have the same surface point type, Ambivalent H/LP, this allows eHiTS to handle all three possible protonation states in a single run.



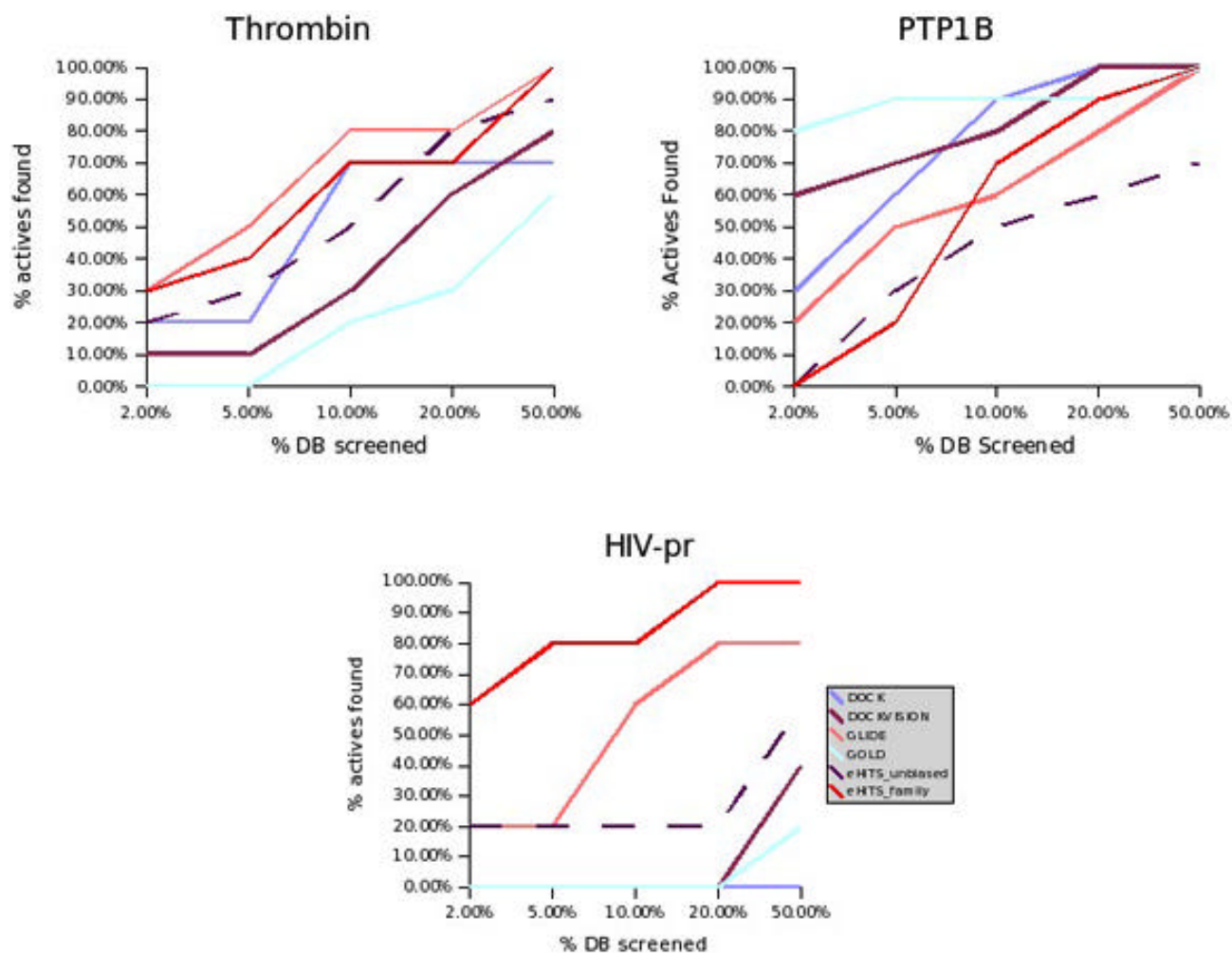
**Fig. (8).** Docking accuracy of eHiTS for both the training set (884 complexes) and the test set (742 complexes), (a) shows results for the top-ranked, best scoring pose, (b) shows the result for the closest pose found in the top 32 reported poses.



**Fig. (9).** Comparison of docking accuracy between the complexes recognized as a member of a family and those that are not. (a) shows the results for the top-ranked pose for those identified as part of a family, (b) shows the results for the singletons or those not part of a family.



**Fig. (10).** Comparison of docking accuracy between the complexes recognized as a member of a family and those that are not. (a) shows the results for the closest pose found for those identified as part of a family, (b) shows the results for the singletons or those not part of a family.



**Fig. (11).** Virtual screening results for three receptor families, thrombin, PTP1b and HIV-Pr.

In general the family-based training does dramatically improve validation results and does improve screening results to varying degrees.

## 9. DISCUSSION

The eHiTS docking program has been shown to be able to reproduce the x-ray poses of bound ligand with extremely high accuracy, across a wide variety of families. The novel approach to the scoring function problem does appear to have some merit in this initial test. Further work and analysis will have to be done to fully explore the limits and capabilities of the new scoring function. However, with the training as described in section 3.1, it should be possible to improve the results even more, and possibly tailor the scoring function to particular types of receptors or ligands as desired by users.

The eHiTS algorithm of docking rigid fragments independently everywhere in a receptor site not only eliminates any seed bias but also allows for the reuse of the docking information in subsequent docking runs to improve the speed of the docking algorithm. In addition the automatic handling of protonation state allows eHiTS to test all possible proto-

nation states of a receptor-ligand pair and thus report the most appropriate form, according to the scoring function.

The eHiTS docking program offers users a different kind of docking experience: full automation, no need for timely preparation of receptor and ligand structures (protonation, partial charges, energy minimization not required), and no compromise in results with highly specific family-trained scoring function. eHiTS is free to academic users.

## REFERENCES

- [1] PDB.org <http://www.rcsb.org/pdb/holdings.html>
- [2] Irwin, J.J. and Shoichet, B.K. (2005) *J. Chem. Inf. Model.*, 45(1), 177-82.
- [3] PubChem <http://pubchem.ncbi.nlm.nih.gov/>
- [4] Goodsell, D.S. and Olson, A.J. (1990) *Proteins Structure Function and Genetics*, 8, 195-202.
- [5] Hart, T.N., Ness, S.R. and Read, R.J. (1997) *Proteins*, (Suppl., 1), 205-209.
- [6] Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R. (1997) *J. Mol. Biol.*, 267, 727-748.
- [7] Westhead, D.R., Clark, D.E. and Murray, C.W. (1997) *J. Comput. Aided Mol. Des.*, 11, 209-228.
- [8] Rarey, M., Kramer, B., Lengauer, T. and Klebe, G.A. (1996) *J. Mol. Biol.*, 261, 470-89.
- [9] DesJarlais, R.L., Sheridan, R.P., Seibel, G. L., Dixon, J. S. and Kuntz, I. D. (1988) *J. Med. Chem.*, 31(4), 722-729.

- [10] Kearsly, S.K., Underwood, D.J., Sheridan, R.P. and Miller, M.D. (1994) *J. Comput. Aided Mol. Des.*, 8, 565-582.
- [11] McGann, M., Almond, H., Nicholls, A., Grant, J.A. and Brown, F. (2003) *Biopolymers*, 68, 76-90.
- [12] Warren, G.L., Andrews, C.W., Capelli, A.M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Senger, S., Tedesco, G., Wall, I.D., Woolven, J.M., Peishoff, C.E. and Head, M.S. (2005) *J. Med. Chem.*, 10, 1021/jm050362n.
- [13] Stahl, M. and Rarey, M. (2001) *J. Med. Chem.*, 44(7), 1035-1042, 10.1021/jm0003992
- [14] Kitchen, D.B., Decornez, H., Furr, J.R. and Bajorath, J. (2004) *Nature Rev. Drug Discov.*, 3, 935-949.
- [15] Cummings, M.D., DesJarlais, R.L., Gibbs, A.C., Mohan, V. and Jaeger, E.P. (2005) *J. Med. Chem.*, 48(4), 962 - 976, 10.1021/jm049798d
- [16] Bursulaya, B.D., Totrov, M., Abagyan, R. and Brooks, C.L. 3rd. (2003) *J. Comput. Aided Mol. Des.*, 17(11), 755-763.
- [17] Perola, E., Walters, W.P. and Charifson, P.S. (2004) *Proteins*, 56(2), 235-249.
- [18] Kellenberger, E., Rodrigo, J., Muller, P. and Rognan, D. (2004) *Proteins*, 57(2), 225-242.
- [19] Trosset, J.-Y. and Acheraga, H.A. (1998) *Proc. Natl. Acad. Sci. USA*, 95, 8011-8015.
- [20] Abagyan, R. and Totrov, R. (1994) *J. Mol. Biol.*, 235, 983-1002.
- [21] Totrov, R. and Abagyan, R. (1997) *Proteins (Suppl. 1)*, 215-220.
- [22] Oshiro, C.M., Kuntz, I.D. and Dixon, J.S. (1995) *J. Comput. Aided Mol. Des.*, 9 (2) 113-130.
- [23] Ewing, T.J.A., Makino, S., Skillman, A.G. and Kuntz, I.D. (2001) *J. Comput. Aid. Mol. Des.*, 15, 411-428.
- [24] Welch, W., Ruppert, J. and Jain, A.N. (1996) *Chem. Biol.*, 3(6), 449-62.
- [25] Simonson, T., Archontis, G. and Karplus, M. (2002) *Acc. Chem. Res.*, 35, 430-437.
- [26] Kollman, P.A. (1993) *Chem. Rev.*, 93, 2395-2417.
- [27] Muegge, I. and Martin, Y.C. (1999) *J. Med. Chem.*, 42, 791-804.
- [28] Muegge, I. (2000) *Perspect. Drug Discov. Des.*, 20, 99-114.
- [29] Muegge, I. (2001) *J. Comput. Chem.*, 22, 418-425.
- [30] Gohlken, H., Hendlich, M. and Klebe, G. (2000) *J. Mol. Biol.*, 295, 337-356.
- [31] DeWitte, R.S. and Shakhnovich, E.I. (1996) *J. Am. Chem. Soc.*, 118, 11733-11744.
- [32] Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P. (1997) *J. Comput. Aid. Mol. Des.*, 11, 425-445.
- [33] Bohm, H.J. (1994) *J. Comput. Aid. Mol. Des.*, 8, 243-256.
- [34] Bohm, H.J. (1997) *J. Comput. Aid. Mol. Des.*, 12, 309-323.
- [35] Rarey, M., Kramer, B., Lengauer, T. and Klebe, G. (1996) *J. Mol. Biol.*, 261, 470-489.
- [36] Wang, R., Liu, L., Lai, L. and Tang, Y. (1998) *J. Mol. Model.*, 4, 379-394.
- [37] Tao, P. and Lai, L. (2001) *J. Comput. Mol. Des.*, 15, 429-446.
- [38] Wang, R., Lai, L. and Wang, S. (2002) *J. Comput. Aided Mol. Des.*, 16, 11-26.
- [39] Rognan, D., Lauemoller, S.L., Holm, A., Buus, S. and Tschinke, V. (1999) *J. Med. Chem.*, 42, 4650-4658.
- [40] Kramer, B., Rarey, M. and Lengauer, T. (1999) *Proteins*, 37, 228-241.
- [41] Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W. and Taylor, R.D. (2003) *Proteins*, 52, 609-623.
- [42] Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J. (1998) *J. Comput. Chem.*, 19, 1639-1662.
- [43] Ewing, T.J.A., Makino, S., Skillman, A.G. and Kuntz, I.D. (2001) *J. Comput. Aided Mol. Des.*, 15, 411-428.
- [44] Ajay and Murcko, M.A. (1995) *J. Med. Chem.*, 38, 4953-4967.
- [45] Gohlke, J. and Klebe, G. (2001) *Curr. Opin. Struct. Biol.*, 11, 231-235.
- [46] Tame, J.R. (1999) *J. Comput. Aided Mol. Des.*, 13, 99-108. 1999.
- [47] Mugge, I. and Rarey, M. (2001) in *Reviews in Computational Chemistry* (Lipkowitz, K.B. and Boyd, D.B., Eds.), 17, 1-60, John Wiley & Sons, New York.
- [48] Halperin, I., Ma, B., Wolfson, H. and Nussinov, R. (2002) *Proteins*, 47, 409-443.
- [49] Wang, R., Lu, Y. and Wang, S. (2004) *J. Med. Chem.*, 46, 2287-2307.
- [50] Miteva, M.A., Lee, W.H., Montes, M.O. and Villoutreix, B.O. (2005) *J. Med. Chem.*, 48, 6012-6022.
- [51] Sotriffer, C. and Klebe, G., (2002) *Farmaco*, 243-251.
- [52] Campbell, S.J., Gold, N.D., Jackson, R.M. and Westhead, D.R. (2003) *Curr. Opin. Struct. Biol.*, 13(3), 389-395.
- [53] Ruppert, J., Welch, W. and Jain, A.N. (1997) *Protein Sci.*, 6, 424-533.
- [54] Verdonk, M.L., Cole, J.C. and Taylor, R. (1999) *J. Mol. Biol.*, 289, 1093-1108.
- [55] Verdonk, M.L., Cole, J.C., Watson, P., Gillet, V. and Willett, P. (2001) *J. Mol. Biol.*, 307, 841-859.
- [56] Bliznyuk, A.A. and Gready, J.E. (1998) *J. Comput. Aided Mol. Des.*, 12(4), 325-333.
- [57] Bliznyuk, A.A. and Gready, J.E. (1999) *J. Comput. Chem.*, 20, 983-988.
- [58] Laurie, A.T. and Jackson, R.M. (2005) *Bioinformatics*, 21(9), 1908-1916.
- [59] Hendlich, M., Pippmann, F. and Barnickel, G. (1997) *J. Mol. Graph. Model.*, 15, 359-363.
- [60] Levitt, D.G. and Banaszak, I.J. (1992) *J. Mol. Graph.*, 10, 229-234.
- [61] Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A. and Thornton, J.M. (2006) *Proteins*, 62, 479-488.
- [62] Peters, K.P., Fauck, J. and Frommel, C. (1996) *J. Mol. Biol.*, 256, 201-213.
- [63] Liang, J., Edelsbrunner, H. and Woodward, C. (1998) *Protein Sci.*, 7, 1884-1897.
- [64] Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) *Adv. Drug Deliv. Rev.*, 23, 3-25.
- [65] Lam, P.Y.S., Jadhav, P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bachelier, L.T., Meek, J.L., Otto, M.J., Rayner, M.M., Wong, Y.N., Chang, C.H., Weber, P.C., Jackson, D.A., Sharpe, T.R. and Erickson-viitanen, S. (1994) *Science*, 263, 380-384.
- [66] Iversen, L. F., Andersen, H. S., Branner, S., Mortensen, S. B., Peters, G. H., Norris, K., Olsen, O. H., Jeppesen, C. B., Lundt, B. F., Ripka, W., Møller, K. B. and Møller, N. P. H. (2000) *J. Biol. Chem.*, 275, 10300-10307.
- [67] Bone, R., Lu, T., Illig, C.R., Soll, R.M. and Spurlino, J.C. (1998) *J. Med. Chem.*, 41, 2068-2075.
- [68] MDDR, [http://www.mdll.com/products/knowledge/drug\\_data\\_report/index.jsp](http://www.mdll.com/products/knowledge/drug_data_report/index.jsp).