# Do Structurally Similar Ligands Bind in a Similar Fashion?

Jonas Boström,*[,†] Anders Hogner,[‡] and Stefan Schmitt[‡]

*Department of Medicinal Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden, and Global DECS Computational Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden*

*Received February 14, 2006*

The scope of the current work is to investigate whether structurally similar ligands bind in a similar fashion by exhaustively analyzing experimental data from the protein database (PDB). The complete PDB was searched for pairs of structurally similar ligands binding to the same biological target. The binding sites of the pairs of proteins complexing structurally similar ligands were found to differ in 83% of the cases. The most recurrent structural change among the pairs involves different water molecule architecture. Side-chain movements are observed in half of the pairs, whereas backbone movements rarely occurred. However, two structurally similar ligands generally confirm a high degree of structural conservation. That is, a majority of the ligand pairs occupy the same region in the binding sites, providing support for the use of shape matching in the drug design process. We allow ourselves to draw general conclusions because our data set consists of ligands with drug-like physicochemical properties complexed to a broad spectrum of different protein classes.

## Introduction

A central theme in drug design is to make small modifications to lead molecules to obtain desired properties. The rationale of this approach is that similar molecules bind in a similar fashion and would, thus, induce the same desired biological effect. The hypothesis runs like a red line through the entire drug design process, from typical early stage approaches (such as assigning similar compounds from an HTS into clusters) to typical mid-stage approaches (such as making modifications to lead compounds to increase affinity and solubility) and all the way through to typical late stage concerns (such as attending to toxicity and metabolism issues). This hypothesis is so entrenched in the drug design process that it is practically universally believed. It is also the assumption when experimentally determined protein structures are used in structure-based design efforts. Nevertheless, and perhaps somewhat paradoxically, the solution of the atomic positions for a new ligand–protein complex frequently leads to unexpected insights compared with those from previously solved complexes. For example, a ligand can upon binding cause a side-chain to assume another rotamer,[1] and it can affect the presence of water molecules[2] as well as causing larger effects, such as backbone movements[3] and even domain movements.[4] Moreover, the ligand itself can adopt different modes of binding such as flipping into so-called reversed-binding modes.[5–8] Because it is no trivial task to predict any of these events, their effect on the outcome of structure-based design efforts may turn out to be detrimental. One simplistic way to circumvent the many difficulties of protein flexibility is to neglect the protein and simply use the shape of the bioactive conformation of a lead compound alone to find novel compounds. This approach has proven successful in a number of cases. For example, the use of the shape-matching program ROCS[9] has recently led to the identification of a set of novel inhibitors of the ZipA-FtsZ protein–protein interaction.[10] Even so, it is clear that having the 3D coordinates of a ligand–protein complex plays an important role in improving success rates in the drug discovery process,[11] and the exploitation of structural data is crucial in improving the accuracy of structure-based as well as ligand-based techniques.

The aim of this study is to investigate whether similar ligands actually do bind in a similar fashion by systematically analyzing pairs of structurally similar ligands binding to identical biological targets. Because the crystallographic protein–ligand database (PDB)[12] offers the most comprehensive and reliable source of information about ligand–protein interactions, the PDB was searched using Reliscript.[13] The occurrence of events such as protein flexibility and ligand binding modes are recorded in detail for these pairs. There are numerous examples of such events in the medicinal chemistry literature. For instance, Teague recently presented an excellent review including several illustrative examples of the effect of protein flexibility upon ligand binding.[14] However, most publications are concerned with a set of ligands binding to a single biological target. This is the first exhaustive experimental survey on the complete PDB. Here, we present the facts and statistics, and challenge the belief that similar molecules bind in a similar fashion. In addition to supporting the use of shape-matching in drug design, the output of this study can give guidance to medicinal chemists in their decisions toward requesting structural data for their compounds.

## Materials and Methods

With the aim to derive an exhaustive data set of pairs of structurally similar ligands complexed to the binding site of the same biological target, structures from the PDB were filtered using Reliscript and a series of in-house programs. Reliscript[13] is a command-line interface that allows access to PDB data and elaborate search methods from within the Python[15] scripting language environment. The steps for retrieving such a data set are illustrated in Figure 1 and described below.

**Retrieval of Drug-Like Ligands from the PDB.** Starting from the Relibase+ November 2004 data-release with 27 887 PDB structures deposited, only X-ray structures with a resolution better than 2.5 Å and not containing the PDB-tag *CAVEAT* were considered. The remaining data set was queried for *relevant ligands*, which were defined as small molecules that would be of interest to a medicinal chemist during the drug discovery

* To whom correspondence should be addressed. Phone: +46 31 706 52 51. Fax: +46 31 776 37 10. E-mail: jonas.bostrom@astrazeneca.com.

† Department of Medicinal Chemistry.

‡ Global DECS Computational Chemistry.

10.1021/jm060167o CCC: $33.50
Published on Web 10/18/2006

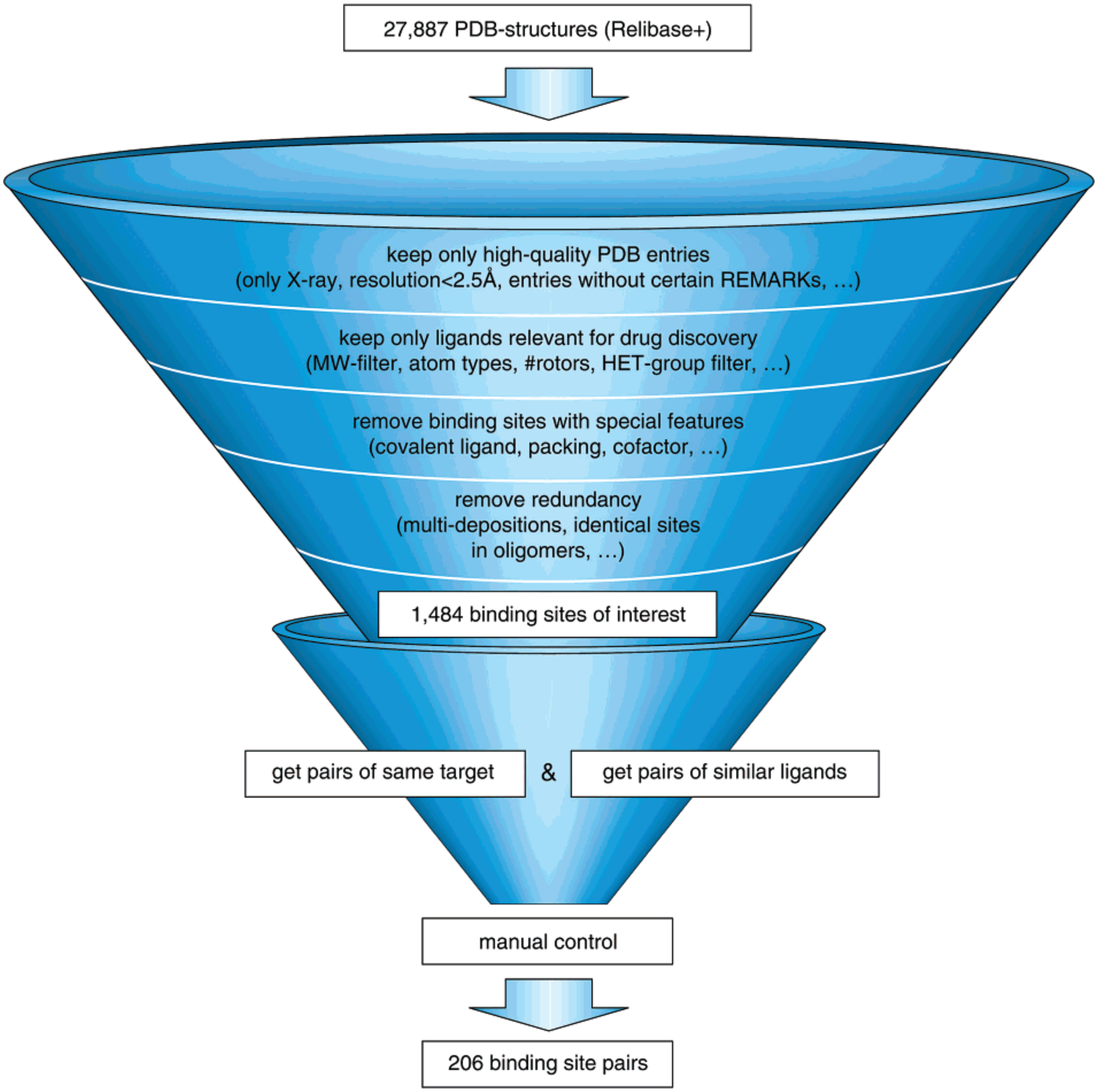

27,887 PDB-structures (Relibase+)
keep only high-quality PDB entries
(only X-ray, resolution<2.5Å, entries without certain REMARKs, …)
keep only ligands relevant for drug discovery
(MW-filter, atom types, #rotors, HET-group filter, …)
remove binding sites with special features
(covalent ligand, packing, cofactor, …)
remove redundancy
(multi-depositions, identical sites
in oligomers, …)
1,484 binding sites of interest
get pairs of same target
&
get pairs of similar ligands
manual control
206 binding site pairs

**Figure 1.** Scheme used for filtering protein structure information from the PDB. The cascade resulted in 206 pairs of binding sites from the same biological targets that complex structurally similar ligands.

process. Identifying such relevant ligands from PDB data required an elaborate filtering cascade accompanied by manual inspection and deselection at various stages because PDB-data also contain small molecules, such as cofactors, crystallization buffers, solvents, and other agents that cannot be straightforwardly distinguished from the true ligand that is in complex with the protein. At the time, Relibase+ registered 161 570 small molecules as ligands which were subject to the empirically defined filter criteria as listed in Table 1. The criteria a−g ensure lead-like properties for small molecules. Previous in-house studies on PDB-data returned a list with 451 PDB-HET group IDs that are typically used for cofactors, sugar molecules, crystallization agents, or other non-lead-like ligands; the criterion h uses this list to remove all Relibase-ligands that contain only these fragments. Criteria i and j exclude ligands with special binding features, whereas k discards further artifacts that are not of interest in the current study. Despite this thorough and systematic processing, the automatic filtering steps for small molecules from PDB structures had to be accompanied by manual inspections to remove additional non-lead-like ligands. A subsequent step removed the redundancy from the remaining data set by applying two additional criteria: for cases of oligomeric structures where the asymmetric unit contains the same ligand in the same protein environment more than once, only one binding site environment was kept, and for PDB-structures representing identical protein−ligand complexes, only the structure with best resolution was kept. The overall selection resulted in a nonredundant data set of 1484 small molecules in 1451 protein−ligand complexes that were of potential relevance for this study.

**Identification of Pairs by Chemical Similarity.** In the next step, pairs of structurally similar ligands were identified from the filtered data set. There are numerous ways to determine if molecules are similar or not. Maximum common substructures (MCSS) provide an intuitively reasonable view of the structural similarity between two molecules and is, thus, well suited to identify structurally similar compounds. Hence, in line with work by Raymond[16] and Barker,[17] we calculated Tanimoto similarities on the basis of the MCSS in a pairwise manner. The *Tanimoto*$_{\text{MCSS}}$ was calculated using the OpenEye OEChem-

**Table 1.** Filter Criteria Used to Retrieve Relevant Ligands from PDB Structural Data as It Is Deposited in Relibase+[a]

| | Criteria to retrieve relevant ligands from the PDB |
|---|---|
| (a) | 80 < molecular weight (Da) < 750 |
| (b) | 10 < number of nonhydrogen atoms < 70 |
| (c) | must not contain atoms of types other than H, C, O, N, F, P, S, Cl, Br, or I |
| (d) | must contain at least one non-carbon/non-hydrogen atom |
| (e) | must not contain two or more phosphorus atoms |
| (f) | must not have more than 10 rotatable bonds |
| (g) | must not be a nucleic acid |
| (h) | must not be composed only from non-lead-like PDB-HET-groups[b] |
| (i) | must not be covalently bound |
| (j) | must not have protein contacts from the crystal packing environment in less than 3 Å distance |
| (k) | must have contacts with protein in less than 7Å distance |

[a] Relevant ligands are defined as small molecules that would be of interest to a medicinal chemist during a drug discovery process. [b] Using a previously derived list with 451 PDB-HET-group IDs for cofactors, sugar molecules, crystallization agents, or other non-lead-like fragments.

toolkit,[18] according to eq 1.

$$Tanimoto_{\mathrm{MCSS}} = \frac{N_{\mathrm{AB}}}{(N_{\mathrm{A}} + N_{\mathrm{B}}) - N_{\mathrm{AB}}} \qquad (1)$$

where $N_A$ and $N_B$ are the number of atoms in molecules A and B, respectively, and $N_{AB}$ is the number of atoms in the MCSS of A and B. The $Tanimoto_{\mathrm{MCSS}}$ can have values between 0 and 1. A value of 1 is obtained if two molecules are identical. By manual inspection, it was established that $Tanimoto_{\mathrm{MCSS}}$ values $\geq 0.8$ disclose structurally similar ligand pairs. This cutoff resulted in 1869 pairs, obtained from the above-described data set of relevant ligands. In the current study, we investigate similar, but not identical, ligands. Consequently, identical ligands ($Tanimoto_{\mathrm{MCSS}} = 1.0$) were removed. Identical pairs of ligands would give valuable information as a positive control. However, only five such examples passed the filtering cascade, which we consider to be too small a number from which to make a statistical inference.

To determine the equivalence of the corresponding binding sites for these ligand pairs, the sequence similarity of the underlying proteins was compared using FASTA[19] as implemented in Relibase+. If the ligand-adjacent protein chains for the two proteins of the underlying ligand pairs have at least 95% sequence identity, then the binding sites were considered to originate from the same biological target. This conservative cutoff still allows for mutations or PDB-structure depositions of the same protein with different *C*- or *N*-terminal lengths. However, for cases where the difference in the sequence of the ligand-adjacent chains falls into the proximity of the binding site, the ligand pairs were manually removed from the data set. Applying the above-described cutoffs results in a data set that contains 397 pairs of similar ligands binding to equivalent protein binding sites. The protein structures for each pair were then superimposed using the align structures by homology option in Sybyl[20] and visually analyzed. This step lead to the removal of additional structures from the data set, such as cases where one binding site contains cofactors or other small molecules in addition to the complexed ligand, whereas the other does not. The reason being that the binding sites for such a pair are not truly comparable. In a few cases, where structure factor files were available, the quality of the X-ray structures was assessed by analyzing electron densities. Accordingly, doubtful structural pairs were also removed. The final manual inspection resulted in 206 binding site pairs with similar ligands. The PDB codes for the binding site pairs obtained in this study are provided in the Supporting Information.

The binding sites for the pairs were subsequently examined for three structural changes that impact the overall binding site environment and, thereby, potentially influence the binding modes for two structurally similar ligands.

(1) Water molecule architecture: if the number of water molecules in the first shell differs, or if they are displaced by more than 2 Å, then the binding site pair is classified as non-identical.

(2) Side-chain rotamers: if the RMSD for all side-chain atoms of at least one amino acid residue within a 5 Å distance from the ligand is greater than 1.0 Å, then the binding site pair is classified as non-identical.

(3) Backbone movements: if the RMSD of at least one backbone heavy atom in three (or more) consecutive amino acids differs by more than 0.5 Å, then the binding site pair is classified as non-identical.

While inspecting the PDB models, it was deemed necessary to flip the glutamine amide functional group 180° for three cases (1i91, 1f06, 1sqn). This was done in order to optimize their hydrogen-bonding network and to make them coherent with their binding site pair partner (1i8z, 3dap, 1a28).

**Calculation of Shape Similarity.** Two types of shape Tanimotos were calculated to assess the shape similarity of the pairs of ligands obtained from the protocol just described and illustrated in Figure 1. ROCS[9] calculations were carried out to obtain shape Tanimoto values, which are a measure of the difference in shape between ligands and are bounded by zero and one. Shape Tanimotos above 0.8 correspond to structures that are visually of very similar shape.

The underlying methodology in ROCS is to compute overlap volumes based on a Gaussian description of molecular shape.[21] Optimal shape Tanimoto values ($Tanimoto_{\mathrm{Shape,Optimized}}$) are found by computing the best alignment between a pair of ligands. An important characteristic of the ROCS method is that it also enables the description of a difference in shape of an arbitrary alignment of a pair of ligands. Such a shape Tanimoto is referred to as an unoptimized shape Tanimoto ($Tanimoto_{\mathrm{Shape,Unoptimized}}$). These were calculated for each ligand pair in the alignment obtained from the superimposed protein structures that are associated with these ligands. This essentially measures how well a ligand retains its 3D position in the binding site upon a minor structural modification, whereas the $Tanimoto_{\mathrm{Shape,Optimized}}$ provides an upper limit on how well the shapes of the bound conformations from two similar ligands match.

## Results and Discussions

**Data Set Composition.** Using the selection criteria described above, we identified a total of 206 binding site pairs. These 412 (206 + 206) protein−ligand structures contain 282 unique proteins and, therefore, 282 unique ligands. Figure 2a−f shows the distribution for standard molecular properties, such as log *P*,[22] molecular weight (MW), the number of rotatable bonds, and the number of hydrogen-bond acceptors/donors, for these ligands. These histograms essentially show the same characteristics as those of the corresponding histograms for compounds with typical lead-like properties.[23]

The data set of 282 unique proteins contains structures from all enzyme classes (EC1−6) as well as structures from various receptor families, such as nuclear hormone receptors and ion-

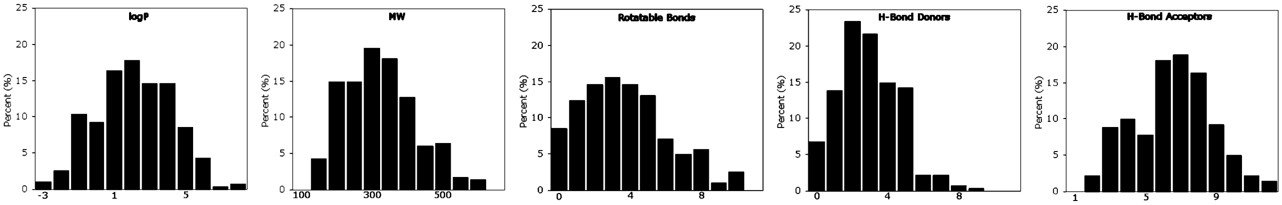


**Figure 2.** (a−f) Distributions of common molecular property parameters for the 282 unique ligands used in the current study.

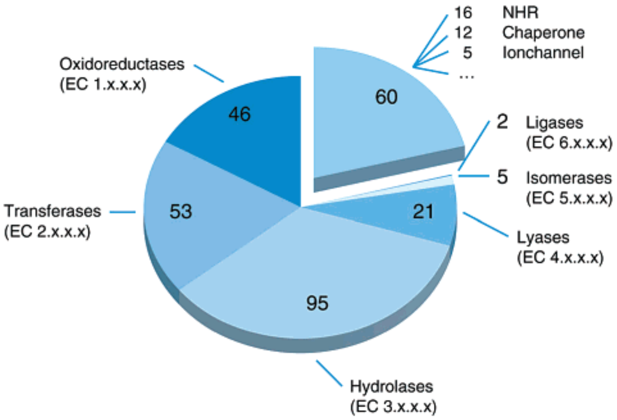


**Figure 3.** Distribution of the 282 unique proteins used in the study, based on the EC classification.

channels, although with different representations. Figure 3 provides a pie-chart based on EC classification[24] and illustrates that all protein classes are reasonably well represented in the data set. It should be noted that Figure 3 reflects a distribution that must be expected when retrieving data from the PDB, that is, an under-representation of ligases (EC6) and isomerases (EC5) with a slight predominance of hydrolases (EC3).[25] Moreover, a great majority of the binding site pairs (96%) have been published by the same crystallographer.

In summary, a systematic and elaborate filtering cascade resulted in a diverse data set of pairs of structurally similar ligands with relevant physicochemical properties complexed to an extensive set of different protein classes. We advocate that such a data set composition allows us to draw general conclusions.

**Analysis of Protein Binding Sites Pairs.** The high-quality X-ray data set of 206 binding site pairs was analyzed on a protein level as well as on a ligand level. The protein binding site pairs were monitored to quantify differences in water molecule architecture, side-chain rotamers, and backbone movements. These events are separately reported below and summarized in Figure 4a−c.

**Water Molecule Architecture.** Water molecules have been shown to be of great importance in the ligand-binding event, both chemically and structurally.[26] For example, water molecules can mediate hydrogen bonds between the ligand and its target protein. Water molecules can also be displaced by the ligand. In this context, it should be noted that the inherent mobility of water molecules make visualization of them using crystallography challenging. Nevertheless, the X-ray structures in the current data set are of high resolution, thus increasing the probability of the water molecules having been correctly identified.

In the vast majority of the 206 complexes, one or more water molecules form hydrogen bonds between the ligand and the protein. When comparing the water molecule architecture within a pair, a difference in the water positions or the number of tightly bound water molecules is observed in as many as 68% of the cases (Figure 4a). The observation of different water molecule architecture cannot be confined to a certain class of proteins but is seen throughout all protein classes.

An example is shown in Figure 5a and b, where a water molecule is mediating a hydrogen bond between backbone atoms in the thrombin protein (PDB code: 1o2g) and a nitrogen in the bound ligand (APC-10302). In the corresponding binding site (PDB code: 1gj4), the corresponding water is displaced by a chlorine substituent in the bound ligand (APC-8696). The indole moiety of the APC-10302 ligand is somewhat differently oriented from that of the matching part in APC-8696 and is approximately 1 Å deeper into the so-called S1 pocket, placing the chlorine well into the cavity produced by the removal of the water molecule. Table 2 shows the molecular structures in 2D as well as the calculated similarity values for the ligand pairs in Figures 5−7, 9, and 10. The analysis shows that the probability of a change in the water molecule architecture within a binding site pair, and thereby the entropic effect on the binding affinity, is well above 50% even for a minor structural modification on the ligand. The prediction of binding affinities by structure-based computational methods (e.g., docking) is, therefore, not only challenged by the scoring function but also by the ability of a particular method to predict the correct water molecule architecture. An accurate prediction of ligand-binding modes requires a tool to individually determine potential water positions for every docked ligand. At present, the majority of docking methods, in their standard form, do not allow the use

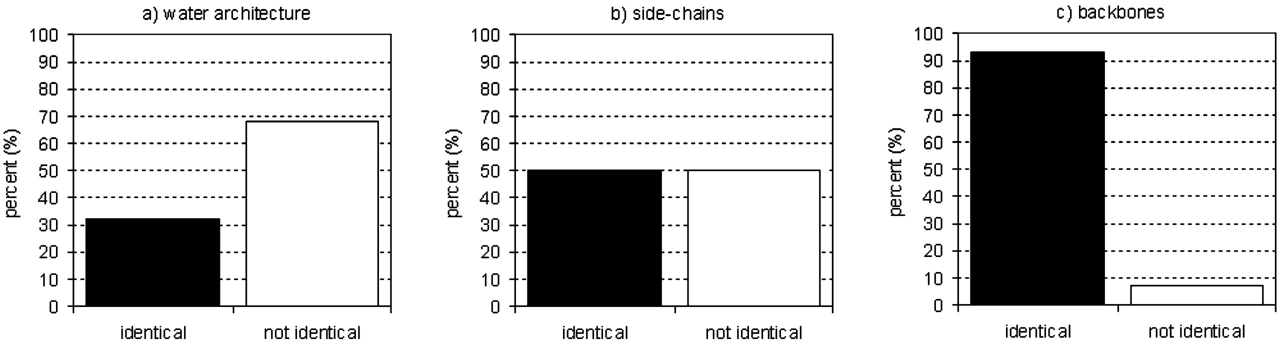


**Figure 4.** (a−c) Distributions of the differences in (a) water molecule architecture, (b) side-chain conformations, and (c) backbone movements for the 206 binding site pairs.

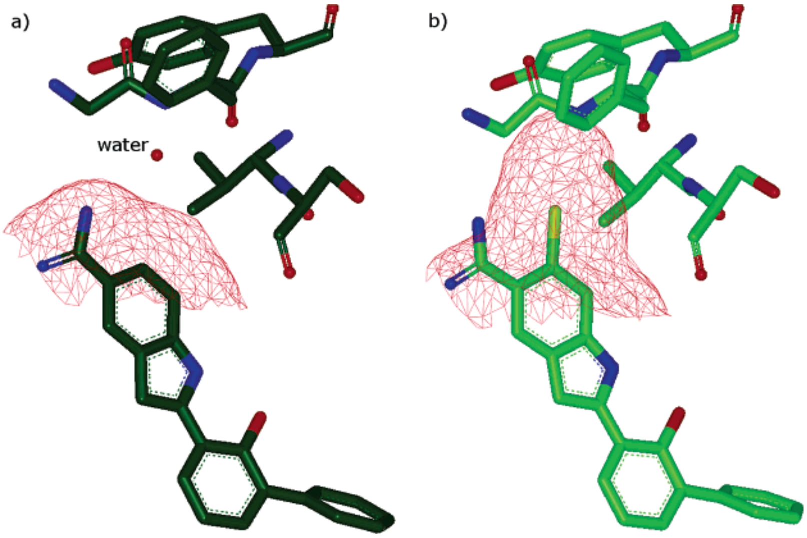


**Figure 5.** Illustration of different water molecule architectures for two structurally similar ligands binding to thrombin. The water in (a) is expelled by a ligand chlorine substituent (light green) in (b). The binding site surface is illustrated with meshed contours colored red.

of different water architectures for different ligands. However, Verdonk et al.[27] recently addressed this issue by implementing a method to score water mediation and displacement in the docking program GOLD.[28] The method allows water molecules to be switched on and off during the docking protocol. In the same spirit, Rarey et al.[29] dealt with this issue in the program FlexX.[30] FlexX can, in a preprocessing phase, calculate favorable positions of water molecules in the binding site and store it in a list. The water molecules are subsequently placed at the precomputed positions, if they can form additional hydrogen bonds to the ligand during docking. These new approaches show potential but need to be further developed to produce an acceptable correlation between calculated and experimental binding affinities. Nevertheless, from the perspective of virtual screens used in hit-to-lead processes, where the aim is to identify a subset of active compounds for a given protein target from a sufficiently large database, omitting water molecules in the calculations is easier to justify because enrichment factors are frequently regarded more important than correctly predicting binding affinities.

**Side-Chain Movements.** Protein conformational flexibility is an important aspect of structure-based drug design. The conformational change of a single side-chain can significantly alter the shape, size, and electrostatics of a binding site. Thus, it can have major consequences for drug design efforts such as *de novo* design and ligand−protein docking.

Side-chain movements are the second most frequent change observed in the current study. The effect of side-chains assuming a different conformation upon or prior to ligand binding is seen in half of the pairs (50%), as shown in Figure 4b. It is not possible to pinpoint a particular protein class or a particular amino acid responsible for this relatively frequent event. That is, there is a roughly equal distribution of side-chain changes among all of the protein classes as well as among the types of amino acids involved.

The crystal structures of the GluR2 ligand-binding core in complex with two structurally similar ligands are shown in Figure 6a and b. In this particular case, the structures of the bound ligands differ by one substituent. The ligand in PDB entry 1mqi has a fluoro substituent (Fluoro-Willardiine), whereas the corresponding ligand in PDB entry 1my3 contains a bromo substituent in the equivalent position (Bromo-Willardiine). To optimally accommodate the specific ligand, the protein side chain of MET:196 assumes two different rotamers in the two binding sites. The contours visualizing the binding site in Figure 6 clearly show the difference in volume and electrostatics for the two binding sites, the binding site accommodating Bromo-Willardiine being largest.

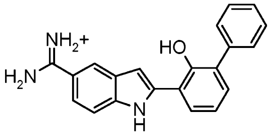
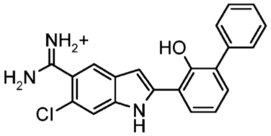
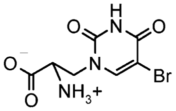
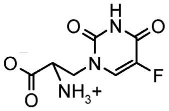
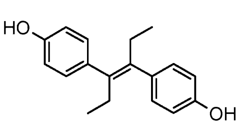
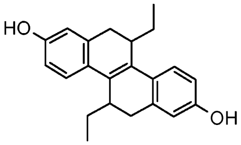
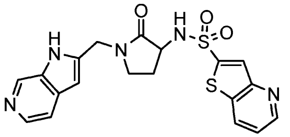
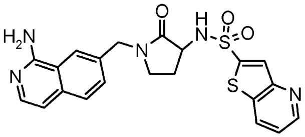
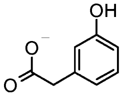
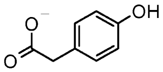
It is apparent from the results obtained in the current study that the use of a single rigid protein structure is in most cases too primitive for accurately docking ligands into proteins. Nevertheless, this is still standard practice in current docking protocols. Although a number of docking algorithms are likely to fit a ligand into a binding site, and possibly produce a reasonable geometry of the bound conformations, success and accuracy of predictions drops dramatically in a more common situation with conformational changes of side-chains. However, structure-based design programs are in a state of rapid development. Progress has already been made in improving the computational methods to accommodate protein flexibility.[31−39] For example, Sherman and co-workers[31] recently published an induced-fit method that accounts for both ligand and receptor flexibility by combining rigid-receptor docking (Glide[40]) with protein structure prediction (Prime[31]) techniques. Correspondingly, Cavasotto and co-workers recently presented a method called the ICM-flexible receptor docking algorithm (IFREDA) to account for protein flexibility in virtual screening.[32]

**Backbone Movements.** Minor structural modifications to a ligand might also cause larger effects on the protein such as changes in the secondary and tertiary structures. These are known as backbone movements and perceived to be the most problematic event to predict in structure-based design efforts.

Backbone movements are observed in no more than 7% of the binding site pairs in our data set. The incidences are again spread over a number protein classes (transferases, nuclear hormone receptors, cocaine anti-bodies, chaperones, and lyases).

Figure 7 shows an example of a backbone movement for the binding site of two structurally similar molecules, diethylstilbestrol (DES) and (*R*,*R*)-5,11-*cis*-diethyl-5,6,11,12 tetrahydrochrysene-2,8-diol (THC), binding to estrogen receptors (ERα) with PDB codes 3erd and 1l2i, respectively. Both ligands act as ERα agonists and, consequently, stabilize the agonist conformation of the receptor. However, when analyzing the binding sites in detail, it is clear that THC induces a backbone movement compared to that of the DES binding site. The backbone movement is induced by the bulky diethyl substituents

**Table 2.** Data for the Ligand-Pairs in Figures 5−7, 9, and 10

| Figure | Protein | PDB codes | | Ligand-Pair Structures | | Ligand-Pair Similarities | | | Protein Movements | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $Tanimoto_{MCSS}$ | $Tanimoto_{Shape,Optimized}$ | $Tanimoto_{Shape,Un\text{-}Optimized}$ | Waters | Side-chains | Backbone |
| **5** | Thrombin | 1o2g | 1gj4 |  |  | 0.96 | 0.96 | 0. 89 | Yes | Yes | No |
| **6** | GluR2 | 1my3 | 1mqi |  |  | 0.88 | 0.99 | 0.93 | No | Yes | No |
| **7** | Estrogen | 3erd | 1l2i |  |  | 0.83 | 0.72 | 0.62 | No | Yes | Yes |
| **9** | Factor Xa | 1f0s | 1f0r |  |  | 0.82 | 0.95 | 0.89 | Yes | Yes | No |
| **10** | Dioxygenase | 3pce | 3pcg |  |  | 0.83 | 0.93 | 0.66 | Yes | No | No |

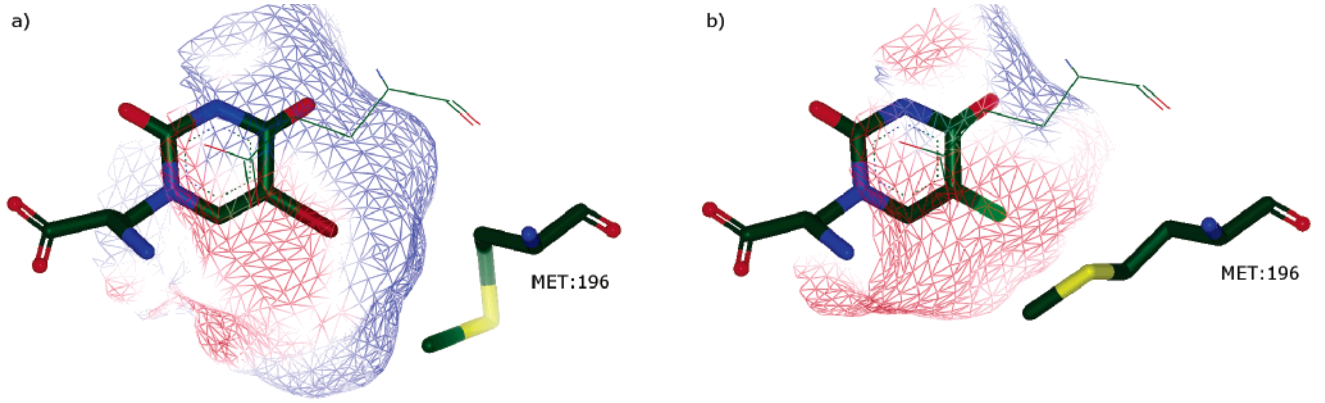


**Figure 6.** Illustration of a side-chain movement for two structurally similar ligands binding to the GluR2 protein. The different ring substituents in Bromo-Willardiine (a) and Fluoro-Willardiine (b) cause side-chain MET:196 to assume two different rotamers. As a result, the shape and electrostatics of the binding site are considerably altered. The binding site surface is displayed by meshed contours colored according to the electrostatic potential.

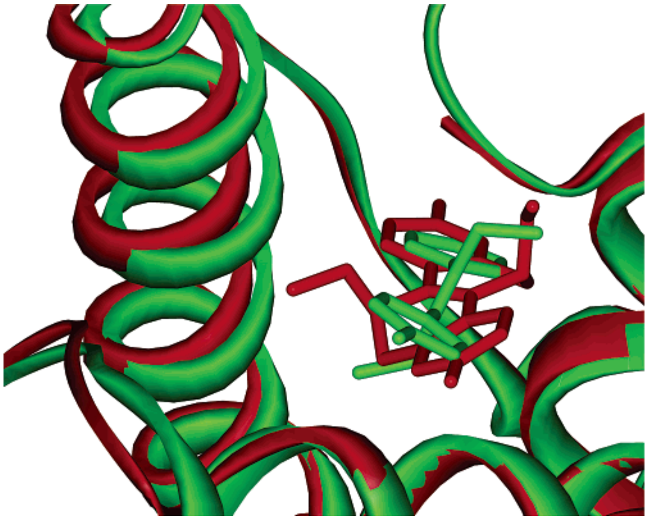


**Figure 7.** Illustration of a backbone movement for two structurally similar ligands binding to estrogen receptors (ERα). One helix in the protein complexing the bulky ligand (red) is shifted to the left compared to the position of the helix of the protein complexing the less bulky ligand (green).

in THC, which are in close proximity to a helix composing the binding site. As a result, it forces the helix to be positioned further away from the location of its counterpart in the DES complex structure (the heavy atom RMSD for three consecutive backbone amino acids differs by 1.0 Å). The shapes and the sizes of the two binding sites differ significantly.

We conclude that backbone movements do occur even for structurally similar molecules, yet it can be seen as an exception. From a drug design perspective, this is fortunate because backbone movements are the most difficult to account for. Nevertheless, incorporating binding site flexibility in proteins will have to go beyond modeling water molecule architectures and side-chain conformations.

**Analysis of the Ligand Pairs. Shape Similarity.** The distribution of optimized shape Tanimoto ($Tanimoto_{\mathrm{Shape,Optimized}}$) values for each pair are displayed in Figure 8. It should be noted that shape Tanimoto values greater than 0.8 are very similar in shape and have a high probability of functional similarity.[10] The results obtained after optimizing the shape overlays with ROCS are encouraging. As many as 186 pairs (90%) show $Tanimoto_{\mathrm{Shape,Optimized}}$ values greater than 0.8. Most ligands would consequently be very highly ranked in a database search using ROCS (provided the correct conformations were in the conformational ensemble), supporting the design strategy of identifying similar compounds by shape matching. The ligand pairs that showed mediocre shape similarity after ROCS alignment can be rationalized by three observations: the ligand X-ray conformations were too different, the ligands had different

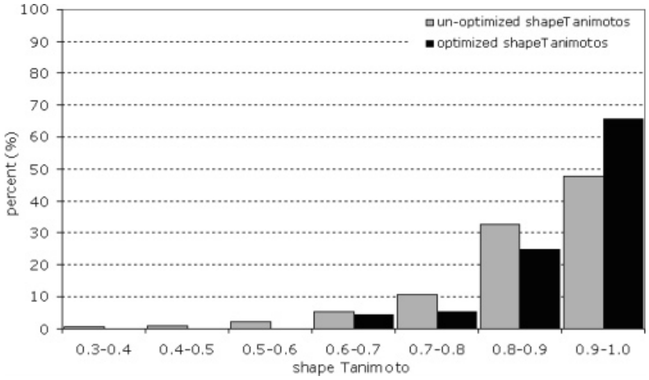


**Figure 8.** Histogram of binned shape Tanimotos for the 206 binding site pairs. A majority of the ligand pairs (90%) show shape Tanimoto values greater than 0.8 after shape optimization (black bars). The shape Tanimotos for the ligand pairs in the orientation obtained after superimposing the protein X-ray structures are generally high, that is, 80% of the pairs show shape Tanimotos above 0.8 (grey bars), indicating that the binding mode as defined by the relative orientation of the ligands in the protein is in most cases conserved.

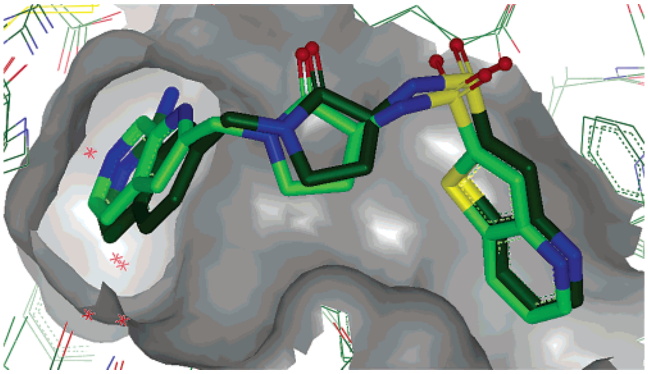


**Figure 9.** Illustration of two structurally similar ligands of high molecular weight binding to a human factor Xa protein. The shape Tanimoto for the fixed alignment is 0.89, meaning that the ligands essentially occupy the same space in the binding site. Hence, the two ligand structures confirm a high degree of structural conservation. The binding site surface is displayed by solid contours colored gray.

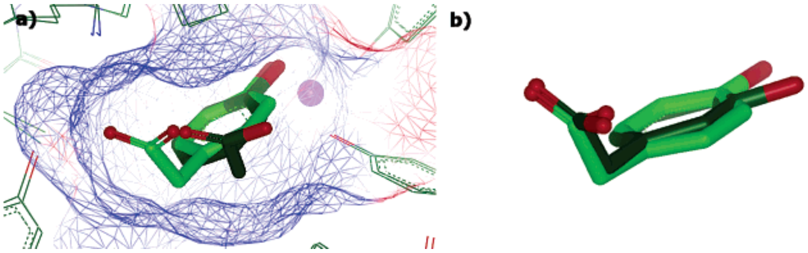


**Figure 10.** (a) Illustration of two structurally similar ligands of low molecular weight binding to a dioxygenase protein. The two carboxylate groups do not occupy the same region in the binding sites. The shape Tanimoto for the fixed alignment is 0.66, and the relative orientation of the common fragment is not conserved. The binding site surface is illustrated with meshed contours colored according to the electrostatic potential. (b) The optimized overlay. The shape Tanimoto is 0.93, and the structural positions of the two compounds are virtually identical.

chirality, and ROCS failed to produce the best possible alignment.

In addition, to investigate if the pairs of similar ligands occupy the same 3D region in the binding site, the unoptimized shape Tanimoto ($Tanimoto_{\mathrm{Shape,Unoptimized}}$) values were also calculated. That is, these shape Tanimotos were calculated for the ligands in the fixed orientation obtained from the alignment of the protein X-ray structures. The histogram in Figure 8 shows that a majority of the ligand pairs occupy the same region in the binding sites. That is, 80% of the pairs show $Tanimoto_{\mathrm{Shape,Unoptimized}}$ values greater than 0.8. This gives emphasis to the fact that structurally similar molecules generally occupy the same region in the binding site. The results are consistent with the recent work by Hare and co-workers.[41] They showed that a majority of kinase ligands that contained related molecular frameworks are found in a single orientation. From a docking perspective, this is an algorithmically useful observation because it allows poses to be discarded where the common framework does not overlap with previously determined experimental structures. Moreover, this concept can also be used in conjunction with docking algorithms based on incremental construction[42,43] because the weakness of such approaches is the need for accurately defining the base fragment.

For a fraction of the pairs (41 out of 206), it is apparent from visual inspection that the overlay is not optimal. These pairs have $Tanimoto_{\mathrm{Shape,Unoptimized}}$ values less than 0.8. A number of physicochemical descriptors were calculated to see if it was possible to identify an intrinsic characteristic associated with these ligand pairs. However, mapping the chemical space of the pairs of ligands shed little light on structural aspects that could potentially be used to discriminate good versus mediocre shape overlays. Only the molecular weight parameter provided some useful insight. If the ligand pairs were classified by molecular weight, it becomes apparent that pairs with MW >370 Da essentially all show high $Tanimoto_{\mathrm{Shape,Unoptimized}}$ values, above the cutoff of 0.8. The pairs of ligands of lower molecular weight (MW <370 Da) have a slightly higher probability of occupying different regions in the binding site. These results are consistent with the notion of molecular complexity.[44,45] The more complex a ligands is, the more stringent the mode of binding should be. Furthermore, it should also be noted that even the pairs with a mediocre overlap still show a significant degree of structural conservation.

The crystal structures of human factor Xa complexed with two structurally similar compounds with MW >370 Da are shown in Figure 9. The structures include different basic groups (the so-called P1 fragments), an aminoisoquinoline group in ligand RPR208815 (PDB entry: 1f0r), and an azaindole group in ligand RPR208707 (PDB entry: 1f0s). The two ligand X-ray structures confirm a high degree of structural conservation as well as the spatial congruence of the aminoisoquinoline and azaindole groups. Consequently, the $Tanimoto_{\mathrm{Shape,Unoptimized}}$ value is high (0.89).

Figure 10a illustrates two ligands of a rather low $Tanimoto_{\mathrm{Shape,Unoptimized}}$ value (0.66) and low molecular weight (MW: 152 Da) complexed with the dioxygenase protein *Pseudomonas putida*. The ligands 3-hydroxy-phenylacetate (MHP) and 4-hydroxy-phenylacetate (PHP) differ only by the position of the hydroxy substituent. Both MHB (PDB entry: 3pce) and PHP (PDB entry: 3pcg) coordinate an iron ($Fe^{3+}$) through its phenolate moiety. As a consequence, the carboxylate groups point into two different regions in the binding sites. Despite this minor structural modification, the positions of the two compounds are not identical. The relative orientation of the ligands is not conserved, and the mediocre shape overlay is reflected in the relatively low $Tanimoto_{\mathrm{Shape,Unoptimized}}$ value. The shape Tanimoto for the optimized overlay is as high as 0.93. The overlay is shown in Figure 10b.

It should be mentioned that no examples of reversed binding modes are found in our data set. The reasons for known examples[5−8] not being present are 3-fold: (i) the ligand−protein coordinates are not available from the PDB; (ii) the ligand pairs did not pass our X-ray resolution filter; and (iii) the ligand pairs did not pass our similarity cut off. Unexpected binding modes certainly exist. But in view of the results presented in the current work, the probability of such an event is very low for structurally similar molecules.

## Conclusions

General conclusions and knowledge about ligand binding modes and protein flexibility are decisive for the drug design process, particularly in the structure-based part of it. We have, therefore, presented the first example of a comprehensive experimental survey on the complete PDB using an elaborate filtering cascade. A diverse set of structurally similar pairs of lead-like ligands spanning a broad range of proteins has been obtained and systematically analyzed.

In general, if two ligands are structurally similar, then the binding mode as defined by the relative orientation in the protein is conserved. That is, a majority of the ligand pairs occupy the same space in the binding sites, in terms of showing shape Tanimoto values greater than 0.8. This is especially true for ligand pairs with molecular weights greater than 370 Da. Practically all pairs show a $Tanimoto_{\mathrm{Shape,Optimized}}$ value greater than 0.8, supporting the design strategy of identifying novel compounds by shape similarity to a known compound.

Nevertheless, two binding sites complexing two structurally similar ligands often have surprisingly different shapes and water architectures. The most frequent structural change involves tightly bound water molecules. Side-chain movements are observed in half of the pairs, whereas backbone movements rarely occur. With the results at hand, it is clear that method development efforts in structure-based drug design should primarily address protein flexibility rather than finding the holy grail of scoring functions.

In conclusion, the probability of a surprise-binding mode is very low. Two structurally similar ligands (e.g., ligands belonging to the same structural series in a drug design project) can safely be assumed to occupy the same 3D position in the binding site. However, there is a significant possibility that minor modifications on a ligand will produce changes in the binding sites that arise from side-chain movements. These changes may give rise to the unexpected structure−activity relationships that are generally seen in drug design projects.

**Acknowledgment.** We thank Drs. Andrew Grant, Peder Svensson, Volker Schnecke, Florian Raubacher, and Jens Sadowski for helpful discussions.

**Supporting Information Available:** The statistics for the 206 binding site pairs used in this study as well as the list of 451 PDB-HET-group IDs for non-lead-like fragments. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Side-Chain Flexibility in Proteins upon Ligand Binding. *Proteins: Struct., Funct., Genet*. **2000**, *39*, 261−268.

(2) Sleigh, S. H.; Seavers, P. R.; Wilkinson, A. J.; Ladbury, J. E.; Tame, J. R. Crystallographic and Calorimetric Analysis of Peptide Binding to OppA Protein. *J. Mol. Biol.* **1999**, *291*, 393−415.

(3) Urzhumtsev, A.; Tete-Favier, F.; Mitschler, A.; Barbanton, J.; Barth, P.; Urzhumtseva, L.; Biellmann, J. F.; Podjarny, A.; Moras, D. A 'Specificity' Pocket Inferred from the Crystal Structures of the Complexes of Aldose Reductase with the Pharmaceutically Important Inhibitors Tolrestat and Sorbinil. *Structure* **1997**, *5*, 601−612.

(4) Armstrong, N.; Gouaux, E. Mechanisms for Activation and Antagonism of an AMPA-Sensitive Glutamate Receptor: Crystal Structures of the GluR2 Ligand Binding Core. *Neuron* **2000**, *1*, 165−181.

(5) Reich, S. H.; Melnick, M.; Davies, J. F., II; Appelt, K.; Lewis, K. K.; Fuhry, M. A.; Pino, M.; Trippe, A. J.; Nguyen, D.; Dawson, H.; Wu, B.; Musick, L.; Kosa, M.; Kahil, D.; Webber, S.; Gehlhaar, D. K.; Andrada, D.; Shetty, B. Protein Structure-Based Design of Potent Orally Bioavailable, Nonpetide Inhibitors of Human Immunodefiency Virus Protease. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3298−3302.

(6) Stoll, V.; Stewart, K. D.; Maring, C. J.; Muchmore, S.; Giranda, V.; Gu, Y.-G. Y.; Wang, G.; Chen, Y.; Sun, M.; Zhao, C.; Kennedy, A. L.; Madigan, D. L.; Xu, Y.; Saldivar, A.; Kati, W.; Laver, G.; Sowin, T.; Sham, H. L.; Greer, J.; Kempf. D. Influenza Neuraminidase Inhibitors: Structure-Based Design of a Novel Inhibitor Series. *Biochemistry* **2003**, *42*, 718−727.

(7) Maignan, S.; Guilloteau, J.-P.; Choi-Sledeski, Y. M.; Becker, M. R.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Molecular Structures of Human Factor Xa Complexed with Ketopiperazine Inhibitors: Preference for a Neutral Group in the S1 Pocket. *J. Med. Chem.* **2003**, *46*, 685−690.

(8) Stubbs, M. T.; Reyda, S.; Dullweber, F.; Möller, M.; Klebe, G.; Dorsch, D.; Mederski, W. W. K. R.; Wurziger, H. pH-Dependent Binding Modes Observed in Trypsin Crystals: Lessons for Structure-Based Drug Design. *ChemBioChem*. **2002**, *2*, 246−249.

(9) *ROCS*, OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com.

(10) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(11) Hardy, L. W.; Malikayil, A. The Impact of Structure-Guided Drug Design on Clinical Agents. *Curr. Drug Discovery* **2003**, *15*, 15−20.

(12) Bernstein, F.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Schimanouchi, T.; Tasumi, M. J. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(13) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Design and Development of a Database for Comprehensive Analysis of Protein−Ligand Interactions. *J. Mol. Biol.* **2003**, *326*, 607−620.

(14) Teague, S. J. Implications of Protein Flexibility for Drug Discovery. *Nat. Rev. Drug. Discovery* **2003**, *7*, 527−541.

(15) Python is an interpreted, interactive, object-oriented programming language. http://www.python.org.

(16) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305−316.

(17) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 503−511.

(18) *OEChem-C++ Theory Manual*, OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com.

(19) Pearson, W. R.; Lipman, D. J. Improved Tools for Biological Sequence Analysis, *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444−2448.

(20) *SYBYL 7.0*, Tripos Inc., 1699 South Hanley Rd., St. Louis, MO, 63144.

(21) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison. A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.

(22) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.

(23) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. I. The Design of Leadlike Combinatorial Libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743−3748.

(24) *Enzyme Nomenclature*; Webb, E. C. Ed.; Academic Press: San Diego, CA, 1992. http://www.chem.qmw.ac.uk/iubmb/enzyme.

(25) Mestres, J. Representativity of Target Families in the Protein Data Bank: Impact for Family-Directed Structure-Based Drug Discovery. *Drug Discovery Today* **2005**, *10*, 1629−1637.

(26) Ladbury, J. E. Just Add Water! The Effect of Water on the Specificity of Protein−Ligand Binding Sites and Its Potential Application to Drug Design. *Chem. Biol.* **1996**, *3*, 973−980.

(27) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling Water Molecules in Protein−Ligand Docking Using GOLD. *J. Med. Chem.* **2005**, *48*, 6504−6515.

(28) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein−Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609−623.

(29) Rarey, M.; Kramer, B.; Lengauer, T. The Particle Concept: Placing Discrete Water Molecules during Protein−Ligand Docking Predictions. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 17−28.

(30) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(31) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534−553.

(32) Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.

(33) Totrov, M.; Abagyan, R. Ab Initio Prediction of Lysozyme-Antibody Complex with 1.6 Å Accuracy. *Nat. Struct. Biol.* **1994**, *1*, 259−263.

(34) Jackson, R. M.; Gabb, H. A.; Sternberg, M. J. Rapid Refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem. *J. Mol. Biol.* **1998**, *276*, 265−285.

(35) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl C. A.; Baker, D. Protein−Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* **2003**, *331*, 281−299.

(36) Zacharias, M. Docking with a Reduced Protein Model Accounting for Side-Chain Flexibility. *Protein Sci.* **2003**, *12*, 1271−1282.

(37) Lorber, D. M.; Udo, M. K.; Shoichet, B. K. Protein−Protein Docking with Multiple Residue Conformations and Residue Substitutions. *Protein Sci.* **2002**, *11*, 1393−1408.

(38) Fernandez-Recio, J.; Totrov, M.; Abagyan, R. ICM-DISCO Docking by Global Energy Optimization with Fully Flexible Side-Chains. *Proteins* **2003**, *52*, 113−117.

(39) Smith, G. R.; Fitzjohn, P. W.; Page, C. S.; Bates, P. A. Incorporation of Flexibility into Rigid-Body Docking: Applications in Rounds 3−5 of CAPRI. *Proteins* **2005**, *60*, 263−268.

(40) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(41) Hare, B. J.; Walters, W. P.; Caron, P. R.; Bemis, G. W. CORES: An Automated Method for Generating Three-Dimensional Models of Protein/Ligand Complexes. *J. Med. Chem.* **2004**, *47*, 4731−4740.

(42) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(43) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499−511.

(44) Hann, M. M.; Leach, R. L.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.

(45) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.

JM060167O