Chem Biol Drug Des 2012; **80:** 182–193

Research Article

doi: 10.1111/j.1747-0285.2012.01396.x

Generation of Receptor Structural Ensembles for Virtual Screening Using Binding Site Shape Analysis and Clustering

David J. Osguthorpe[1], Woody Sherman[2] and Arnold T. Hagler[1,3,*]

[1]*Shifa Biomedical, 1 Great Valley Parkway, Suite 8, Malvern, PA 19355, USA*

[2]*Schrödinger Inc., 120 West 45th Street, 17th Floor, New York, NY 10036, USA*

[3]*Department of Chemistry, University of Massachusetts, 701 Lederle Graduate Research Tower, 710 North Pleasant Street, Amherst, MA 01003-9336, USA*

**Corresponding author: Arnold T. Hagler, athagler@gmail.com*

Accounting for protein flexibility is an essential yet challenging component of structure-based virtual screening. Whereas an ideal approach would account for full protein and ligand flexibility during the virtual screening process, this is currently intractable using available computational resources. An alternative is ensemble docking, where calculations are performed on a set of individual rigid receptor conformations and the results combined. The primary challenge associated with this approach is the choice of receptor structures to use for the docking calculations. In this work, we show that selection of a small set of structures based on clustering on binding site volume overlaps provides an efficient and effective way to account for protein flexibility in virtual screening. We first apply the method to crystal structures of cyclin-dependent kinase 2 and HIV protease and show that virtual screening for ensembles of four cluster representative structures yields consistently high enrichments and diverse actives. We then apply the method to a structural ensemble of the androgen receptor generated with molecular dynamics and obtain results that are in agreement with those from the crystal structures of cyclin-dependent kinase 2 and HIV protease. This work provides a step forward in the incorporation of protein flexibility into structure-based virtual screening.

Key words: binding site volume, clustering, docking, molecular dynamics

Received 2 January 2012, revised 10 March 2012 and accepted for publication 9 April 2012

Virtual screening is an important part of computer-aided drug design, with many reviews (1–3) and successful applications reported (4–7) in the literature. Although ligand-based methods have been shown to yield high database enrichments in virtual screening (8–10), use of structural information, when available, provides the promise of high enrichments of hits that are not biased by knowledge of existing active compounds. Indeed, structure-based virtual screening has been applied successfully in many instances where novel compounds unrelated to existing known actives were found (11–16). However, docking tools have been most heavily validated in the context of a rigid receptor model for both pose prediction (17–22) and virtual screening (23,24), which helps reduce the conformational sampling space and makes the calculations more tractable from a computational perspective. Speed and throughput constraints are particularly important in virtual screening, considering that calculations are often performed on databases of millions of compounds and turnaround time is expected on the order of days to weeks for the computational approach to have a significant impact in the discovery process.

Although successful efforts have been made to account for protein flexibility in pose prediction (25–28) and structure-based virtual screening (28–31), additional work is needed to develop a standardized protocol that can consistently add value over the standard rigid receptor approach. Rueda *et al.* (32) have focused on the same problem we are attempting to address in this work, namely, how to select an appropriate structural ensemble for virtual screening. Through a large-scale analysis of many X-ray structures, the authors determined that selecting the structures with the largest binding sites produced the highest enrichments (32). We aimed to expand on that general direction by including a shape-based description of the binding site and extending the method to molecular dynamics (MD) ensembles.

The numerous ligand-bound protein structures determined by crystallography offer an implicit sampling of protein flexibility. In each of these structures, the protein adapts to a greater or lesser extent to accommodate the ligand. Although sampling of the receptor conformations is limited to large extent by the diversity explored by the co-crystallized ligands, often this is considerable even with a few diverse ligands. Furthermore, a growing number of targets have many crystal structures, with some on the order of 100 or more complexes. Thus, we can take advantage of diverse structures for docking to account in part for flexibility. Here, we explore this concept by introducing a structure-based virtual screening protocol that uses binding site characterization and clustering of structures to

determine a structural ensemble for ligand docking. First, the binding site of each receptor structure is characterized and volume overlaps are computed between all pairs of binding sites. Then, hierarchical clustering is performed to obtain a set of structures with shape-diverse binding sites. We use crystal structures of HIV-1 protease and cyclin-dependent kinase 2 (CDK2) as model systems to show that it is possible to achieve good enrichment of active compounds from a database by combining the top hits from docking into several diverse crystal structures. The results are found to be consistently on par with the best single structure and better than the average results from all structures in the ensemble. Furthermore, the choice of the single receptor conformation that will produce the best enrichments is not known in advance of the screen, highlighting the value of the ensemble approach to yield consistently results on par with the best single structure.

In cases where sufficient diversity from crystal structures is not available, MD simulations also provide for a way to account for protein flexibility by following molecular trajectories as determined by the equations of motion (33–37). Molecular dynamics could conceivably be used to complement or replace the use of crystal structures, especially in the cases where only a few to none have been determined. Molecular dynamics applications have historically focused on understanding general protein movements and the associated biological implications (38–43). However, results have also been presented where MD has been integrated into the virtual screening workflow (44–46), although defining a robust protocol for using MD to represent structural diversity in structure-based drug design has proven to be challenging (47–50). Although some of this may be a result of the computational costs associated with MD, many important scientific questions still remain to be addressed regarding the role of MD in computer-aided drug design. For example, an MD simulation can generate thousands or millions of structures (i.e., frames), but it is not clear which, if any, of those structures are best to use for docking. To this end, we also apply the same binding site shape-based clustering protocol to a set of structures obtained from the trajectory of a crystal structure of the androgen receptor (AR).

In this work, we first study crystal structures of HIV-1 protease and CDK2 as model systems and show that enhanced enrichments are obtained by combining the top hits from docking into several diverse crystal structures. Then, we apply this method to the AR using a conformational ensemble generated by MD and show that it is possible to use MD structures to improve the database enrichment, both in terms of total hits and diversity of the retrieved actives, relative to the use of a single crystallographic receptor conformation. The protocol presented here can be fully automated with no manual decisions and can be applied for virtual screening in other systems where protein flexibility is thought to be important.

Methods

Overview

The objective of the work presented here was to reduce a large ensemble of receptor conformations to a small subset that can be used in virtual screening to improve database enrichment and diversity of the retrieved actives compared with screening a single target conformation. For the work presented here, we chose to look at ensembles of four members without any optimization of the ensemble size. This number was chosen as a balance between expanding structural diversity and keeping computational costs low. The number is further justified based on previous work that showed ensembles between three and five consistently improved virtual screening enrichments (32). The protocol involves first generating a large set of receptor conformations (we explore independent applications of both x-ray crystallography and MD, although other approaches may work). The receptor conformations are then analyzed and clustered based on the binding site shape. Finally, a shape-diverse subset of binding sites is subjected to docking using a database of decoys and known actives. The results are aggregated and compared with the performance of a single receptor model. Each of the stages is described in detail below.

Binding site clustering

The binding site shape for each structure was computed with SiteMap (51,52) after the structures were superimposed using Cα atoms of receptor residues within 8 Å of the ligand. Then, the volume overlap of the binding site of each pair of structures was computed, and a normalized volume overlap matrix was created. Finally, hierarchical clustering was applied to the volume overlap matrix and the centroid of each cluster was chosen as the cluster representative. Histograms of all pairwise volume overlaps for each target are shown in Figure S4.

SiteMap characterizes a binding site using a grid-based approach where properties are computed at each grid point and rules are used to determine which grid points can be combined to create a binding site. The normalized volume overlap between two SiteMap binding sites is then computed using eqn 1:

$$V_{\text{Normalized Overlap}} = \left(V_{\text{Overlap AB}}/V_A\right) \times \left(V_A/V_{(A+B)}\right) + \left(V_{\text{Overlap AB}}/V_B\right) \times \left(V_B/V_{(A+B)}\right), \quad (1)$$

where $V_{\text{Overlap AB}}$ is the overlap volume between the SiteMap volumes for binding sites A and B, V_A is the volume for binding site A, V_B is the volume for binding site B, and $V_{(A+B)}$ is the sum of the volumes for binding sites A and B. The overlap is computed with the phase_volCalc utility from Schrödinger. Two binding site volumes that perfectly overlap have a score of 1.0 (the overlap volume is normalized by the sum of the two volumes), whereas completely non-overlapping volumes have a score of 0.0.

The pairwise volume overlap computation between all structures produces a similarity matrix, which is then run through the Schrödinger cluster matrix utility to perform hierarchical clustering using average linking. The representative structure for a cluster is determined as the structure whose sum of similarities to all other structures of the cluster is the largest (i.e., the cluster centroid). These cluster representatives form the set of structures to be used for ensemble docking. For the purposes of this study, we set the clustering level to four and used the representative structures from each cluster for the docking studies. Although either more or fewer structures could have

been used, our objective was to develop a practical protocol that would not add excessive computational requirements to the standard single-target virtual screening procedure and still be able to account for structural variability of the target. Structural ensembles of this size have previously been shown to improve virtual screening enrichments consistently (32).

Crystal structure validation systems

Two well-known systems, HIV protease and CDK2, that have extensive structural data available for a variety of ligand complexes were chosen as test cases to study the protocol described above. A set of 135 and 92 structures were selected from the PDB database[a] for HIV protease and CDK2, respectively (see Appendix S1 in Supporting Information for the PDB IDs of all structures used in this work). Each of these protein structures was prepared using the Schrödinger Protein Preparation Wizard, which adds hydrogens, builds side chains and loops with missing atom, determines optimal protonation states for ionizable residues, optimizes the hydrogen bonding network, and performs a constrained geometry minimization of the system.

Molecular dynamics simulations

Molecular dynamics simulations of the AR were performed to assess the effectiveness of using MD to generate the ensemble of structures for docking. For this purpose, we used the trajectory from a simulation of a complex of AR with its physiological ligand dihydrotestosterone (DHT) carried out previously (53). The 1T63 structure of AR bound to DHT and a coactivator peptide from the Fletterick laboratory was used as the basis for the simulation (54). Structures were extracted every 5 ps from a 1-ns portion of the trajectory (the frames from 4 to 5 ns) for a total of 200 structures. These simulations were run with the Gromacs program (36,41) in constant particle, volume, and temperature (NVT) ensemble at 306° K in an explicit solvent box of SPC water molecules extending at least 6.0 Å in all directions from the solute atoms. Longer simulations with replica exchange MD to enhance sampling show that increased binding site diversity can be obtained but that the 5-ns simulations described here provide a reasonable amount of sampling at a relatively small computational expense (63).

Docking

The Glide SP algorithm was used for docking (17,23). This algorithm invokes a series of hierarchical filters developed to optimize docking accuracy and computational efficiency. The receptor site is treated rigidly and is represented by a set of fields of progressively accurate scoring on a precomputed grid. The prescreen leads to a much smaller number of poses that are scored with a more accurate energy model. The resulting ligand poses are then minimized in the field of the receptor using the OPLS_2001 force field with a distance-dependent dielectric. The final step is a Monte Carlo search of the nearby ligand torsional states of 3–6 of the lowest energy poses. Because of the lack of explicit solvent, the molecular mechanics energy inadequately ranks the poses and hence the resulting poses are rescored with a composite empirical score called Emodel, which includes molecular mechanics interactions terms, a modified version of ChemScore (55), and a ligand strain energy term (17,23).

Database enrichment and diversity assessment

The enrichment factor (EF) was computed using eqn 2:

$$\mathrm{EF} = F_a / F_d, \tag{2}$$

where F_a is the fraction of actives found and F_d is the fraction of the database sampled. In this study, we have calculated EFs for the actives found in the top scoring 1% and 4% of the total compounds docked [denoted EF (1%) and EF (4%), respectively]. The top hits from the four structures in the ensemble are combined to get a total of 4% by only considering the unique compounds in the top 1% of each individual target screen, ensuring that a full 4% of the entire set is obtained. Although many other enrichment metrics exist, the early enrichment tends to be the most meaningful for real-world virtual screening campaigns where only a small early fraction of the database will be experimentally tested. However, for completeness we have included receiver operating characteristic curves for all individual targets used in this study (Figure S5) and the ensemble for each target (Figure S6), which allow for the computation of enrichment values at other cutoffs.

The diversity of ligands selected was computed by performing hierarchical clustering on a Tanimoto[b] similarity matrix using MACCs fingerprints (56) as implemented in the cheminformatics package Canvas[c]. A clustering level of 0.7 similarity was chosen based on visual inspection of all actives in the ligand dataset for each target, which roughly separated the active ligands into different 'chemotypes.' Thus, we used this metric to determine the number of 'chemotypes' retrieved from each screen, thereby providing an estimate of the diversity of the hits retrieved. An alternative diversity analysis procedure was also applied based on the maximum fingerprint distance between all retrieved compounds, as described previously (57). However, it produced qualitatively the same results as the chemotype approach and therefore only the chemotype analysis is presented here.

Ligand datasets

The Directory of Useful Decoys (DUD) docking set (58) from the Shoichet group was used for the active and decoy ligands. This set included 52 unique active ligands and 1884 decoys for HIV protease and 49 unique active ligands and 1778 decoys for CDK2. For AR, a set of 73 active ligands and 2627 decoys were used for docking. The number of 'chemotypes' contained in all the actives, found from ligand clustering as described above, was 13, 26, and 9 for the HIV, CDK2, and AR actives, respectively.

Results

To test the binding site shape-based clustering method, we studied HIV protease and CDK2 crystal structures and AR MD snapshots, as described in the Methods section. We used four representative structures from clustering as a balance between increasing diversity and minimizing the computational expense, as the virtual screening calculations scale linearly with the number of receptor structures. Although the exact number of clusters could be optimized, our focus here is on determining whether an automated protocol can be

established to find an ensemble of structures that increase virtual screening enrichment and/or the diversity of the actives retrieved.

HIV protease binding site analysis

The hierarchical clustering of the HIV protease crystal structures resulted in two large clusters and two singletons (on the far right of the clustering dendrogram in Figure S1). Figure 1 shows the ligand binding site shapes of the four cluster representatives, with 1C and 1D representing the two singletons (PDB IDs 1XL2 and 3AID). The structures in Figure 1A,B show the representative structures from the larger clusters and display quite similar shape for most of the binding site. As discussed below, both 1XL2 and 3AID have additional volume in the P2′ pocket (circled in yellow) that projects into the flap loop region of one half of the protease dimer between the strands labeled A′ and B′ in the figure. Note that in the case of 1XL2, there is additional volume in front of this volume under the flaps (circled in blue) that partially hides this additional volume.

To quantify these binding site shape comparisons, we calculated the normalized overlap volumes (Table 1) for the four cluster representatives. This overlap matrix confirms the significant difference in shape between the first two clusters and the two singletons, cluster representatives 3 and 4. Thus, the overlap volume between structures 1 and 2 is 0.56, whereas between these two structures and the singletons 3 and 4, the overlap is approximately 0.35–0.40. In addition, the two outliers are seen to be as different from each other as from the two large clusters. Figure 2 shows an overlay of the two singleton structures and the structure of cluster representative 1. Both of the singleton structures differ from the larger clusters primarily in the flap region.

The two singleton structures contain unique scaffolds for HIV protease inhibitors, shown in Figure 3. The 1XL2 structure contains a ligand based on a pyrrolidinemethaneamine scaffold that adopts a unique and unexpected binding mode (59). The other singleton, PDB ID 3AID, was based on an aminimide peptide isothere (60). In these two cases, the binding mode of the ligand places a phenyl group into a pocket that only these two ligands occupy (see Figure 2). None of the ligands in the other 133 HIV protease structures occupies this space in the binding pocket. The rationale for the significant difference between cluster representatives 3 and 4 is shown in Figure 3. The phenyl group of the pyrrolidine ligand of 1XL2 (colored pink in Figure 3) causes a significant distortion in the flap region of chain A. (The flap region comprises residues 35–57 of chains A and B for HIV protease). No other HIV protease ligand has a bulky group in this position, and the 1XL2 crystal structure is the only one of all 135 structures with this flap distortion.

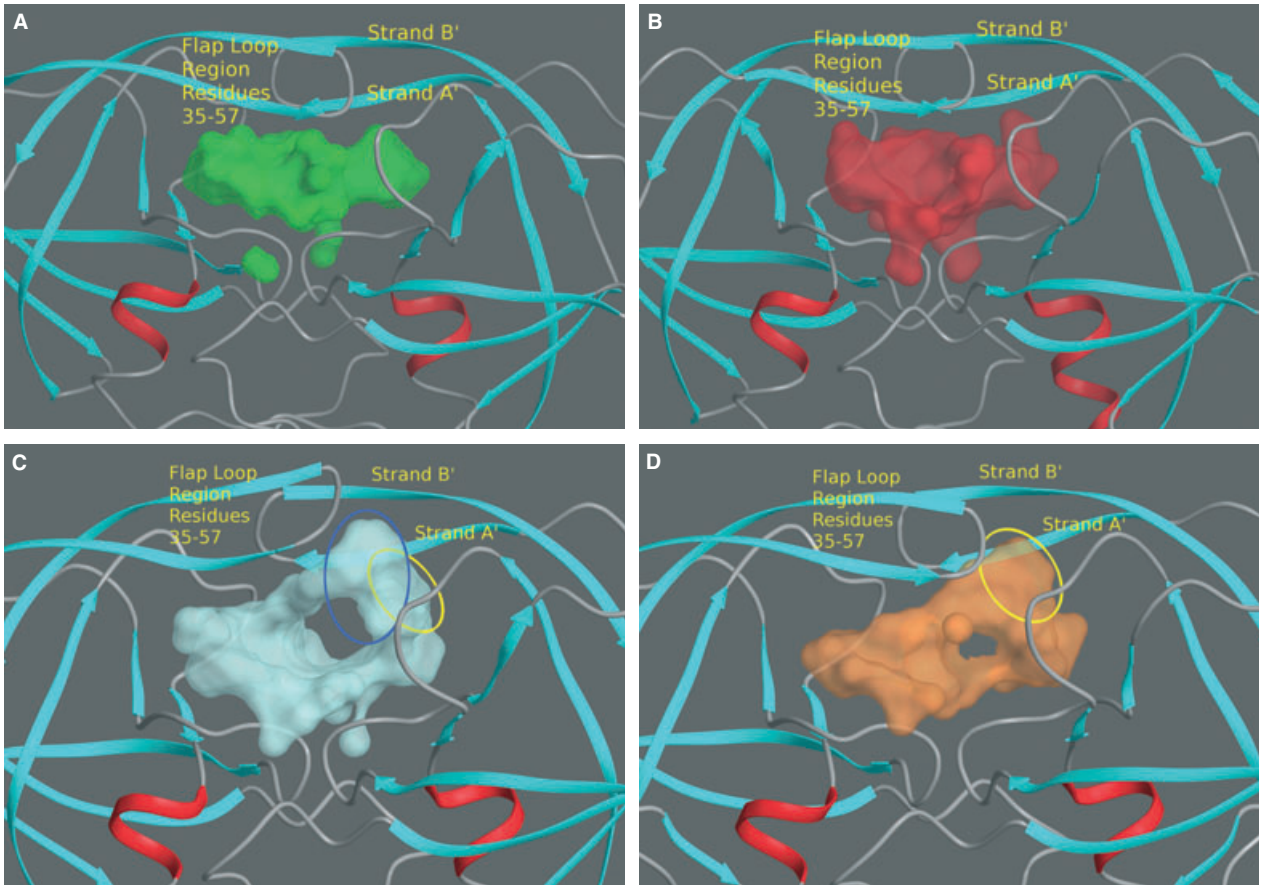

Figure 1: Ligand binding site shapes for the four cluster representatives for HIV protease, PDB IDs 1EBZ (A), 1HVS (B), 1XL2 (C), and 3AID (D).

Table 1: Normalized overlap volume matrix of the four cluster representatives for HIV protease

	1EBZ	1HVS	1XL2	3AID
1EBZ	1.00			
1HVS	0.56	1.00		
1XL2	0.35	0.38	1.00	
3AID	0.39	0.40	0.37	1.00

Specker *et al.* (59) generated a number of new ligands in addition to that in the crystal structure 1XL2. In particular, the crystal structure of a ligand based on an aminohydroxysulfone skeleton bound to HIV protease was determined (PDB ID 1XL5). This structure was also part of our clustering analysis and is found in one of the main two clusters. Specker *et al.* themselves noted that the binding of this ligand was closer to the more expected HIV protease inhibitor binding modes and unlike the pyrrolidine ligand with its novel binding mode. The shape clustering technique definitively captures this difference.

HIV protease virtual screening

Table 2 shows the results for docking the active and decoy dataset into the four cluster representatives, with the enrichment of actives shown for both the individual structures and the ensemble. This table shows the results for the top ranking 1% and 4% of total ligands; the latter allows us to compare ensemble docking using four diverse structures with results that one would achieve by employing a single structure. The bottom row of the table shows the enrichment achieved for the ensemble, which combines the unique hits from the top 1% of each individual structure. As can be seen, the ensemble derived from diverse ligand binding shapes recovers almost 30% more actives than the average obtained from a single crystal structure (13.0 versus 10.3). Whereas the best single structure (from cluster 2) outperforms the ensemble method, the

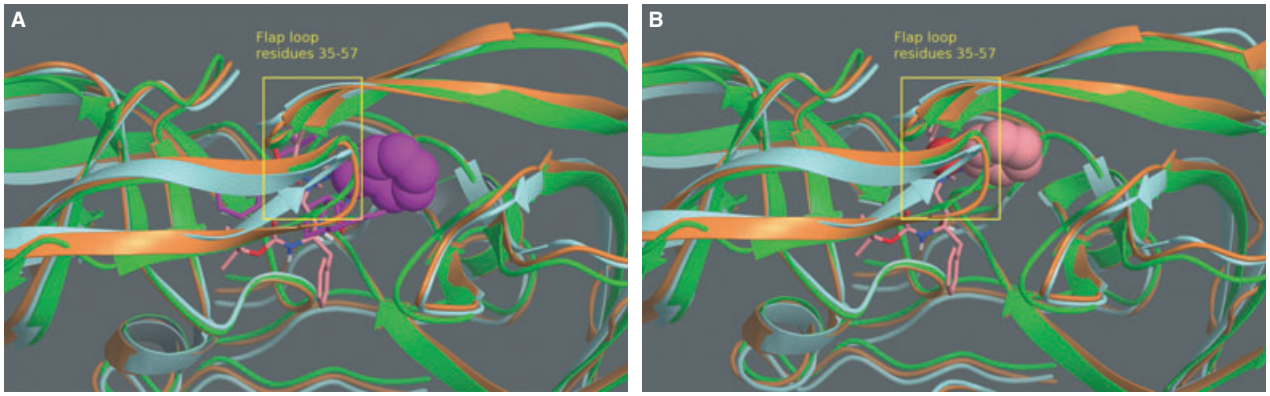

Figure 2: The unique phenyl group binding in the flap of the 1XL2 (A) and 3AID (B) structures is shown in purple and pink space filling representations, respectively. None of the other 133 HIV ligands studied here fills this space. The ligands in cluster representatives 3 and 4 are shown in tube representation with magenta carbons and pink carbons, respectively. This view is from the top relative to Figure 4. The ribbon structures for representatives 1, 3, and 4 are colored green, cyan, and orange, respectively. Representative structure 2 is not shown because it is similar to structure 1.

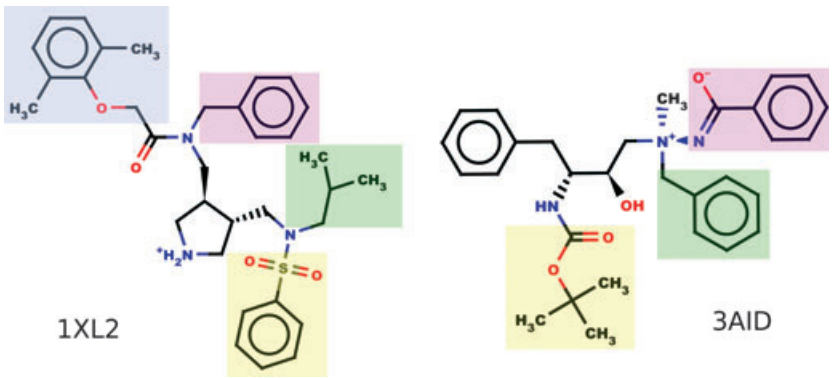

Figure 3: Ligands from outlier cluster structure representatives 3 (1XL2) and 4 (3AID). The shaded groups show those parts of the ligands that are in similar positions in the binding site, with similar shades being in similar positions. The group shaded blue in 1XL2 causes additional structural distortion for this ligand.

Table 2: HIV protease active and decoy ligand docking: actives recovered, enrichment factors, and chemotype count for each structure at the 1% and 4% level

	Count of active ligands		Enrichment factor		Chemotype count	
Cluster representative	Top 1%	Top 4%	EF (1%)	EF (4%)	Top 1%	Top 4%
1EBZ	6 (6)	10	11.5	4.7	3	5
1HVS	10 (5)	16	18.8	7.5	5	6
1XL2	3 (0)	6	5.8	2.8	2	4
3AID	7 (2)	9	13.5	4.2	3	5
Ensemble		13		6.1		6

The unique ligand count is shown in parentheses (i.e., the ligands of the current structure that are not found in any of the previous structures). Chemotype count is the number of clusters represented by the set of actives recovered.

process of choosing a single best target structure is challenging and has only been validated in a limited number of cases (61). These results for HIV protease are encouraging, as the receptor ensemble selection method proposed here results in improved enrichments without an unreasonable increase in computational expense. It is worth noting that the two singletons from the clustering produce the lowest enrichments, as might be expected given that they are the only members of a cluster. However, including diverse structures like this is important because it could be needed to find diverse compounds of interest. For example, in the case here, the overall retrieval is lower with the two singletons, but unique compounds are retrieved that would not have been found otherwise. A standard enrichment metric is not capable of picking up nuances such as this, but it is clearly important for drug discovery when diverse chemotypes are sought.

Diversity of active compounds retrieved

As noted above, not only do we seek to optimize the yield of actives recovered in docking (i.e., enrichment), but we also seek to optimize the 'quality' of the actives we obtain. That is, we would like to recover actives that contain a diverse set of scaffolds and functional groups rather than retrieving a large number of highly related compounds. To give a rough measure of the 'span of molecular space' sampled by the different protocols, we have calculated the chemotype count achieved in the sets of actives (see Methods for description of this metric). As seen in Table 4, the diversity obtained from the ensemble, as assessed by the chemotype count, is as good as the best single-target value. In summary, for HIV protease we achieve better enrichment on average, from using four shape-diverse binding sites for docking over choosing all compounds from a single structure. In addition, the diversity of actives retrieved is on par with the best single structure, which is significant given that the best single-target diversity cannot be determined before the screen is conducted, whereas the shape-based binding site diversity method presented here is deterministic and can be applied without any information about the results.

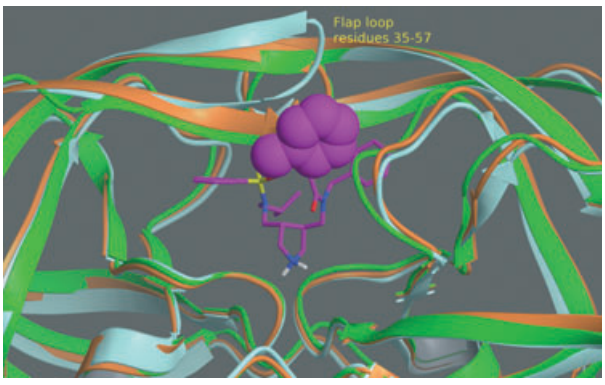

Figure 4: Distortion of the flap for cluster representative 3 (cyan-colored cartoon). The green-colored cartoon shows the backbone structure of cluster representative 1 (PDB ID 1EC3). The orange-colored cartoon is that for cluster representative 4 (PDB ID 3AID). This other singleton cluster only shows a slight distortion in the protein structure. The ligand for cluster representative 3 is shown with magenta carbons.

CDK2 binding site analysis

The clustering of CDK2 binding sites produces a more uniform distribution than we saw with HIV protease (see clustering dendrogram in Figure S2). There are 13, 27, 38, and 14 members in clusters 1 through 4, respectively. Figure 4 shows the ligand binding sites for the four CDK2 cluster representatives. These four shapes each have characteristic features different from the others, as one would expect from the clustering. Most of the changes in binding site shape for these four structures are in the region where the phosphate moieties of ATP bind, although the binding site shape changes each involve a different area of this region.

As with HIV, we calculated the normalized overlap volume matrix for these four cluster representatives to quantify the overlap of these shapes (Table 3). The results reveal a set of fairly small overlap volumes ranging from 0.38 to 0.52. As discussed in the Meth-

Table 3: Normalized overlap volume matrix of the four cluster representatives for cyclin-dependent kinase 2

	1OI9	2BHE	1GZ8	1W0X
1OI9	1.00			
2BHE	0.46	1.00		
1GZ8	0.40	0.38	1.00	
1W0X	0.49	0.52	0.50	1.00

ods section, this indicates that these sites are indeed substantially different and that a large portion of the shape of each cluster representative is sampling protein environments that are not observed in the other structures.

CDK2 virtual screening

In Table 4, we show the results for docking the 1827 active and decoy ligands into the four cluster representative structures shown in Figure 5. Again, we select the top ranked 1% (18) of the hits to each of the four structures, and compare the results of this set of 72 hits with the results we would obtain if we chose the top ranked 72 hits (4%) from docking into any single structure.

The consequences of docking into a small diverse set of CDK2 structures as opposed to a single target structure may be seen by comparing the merged results with those from the single structure docking in Table 4. As with HIV protease, combining the hits from the top ranked 1% of the compounds obtained by docking into each of the four diverse CDK2 structures is superior, on average, to taking the same number of compounds (the top 4%) from docking into a single CDK2 structure (18 hits in top 4% for former versus an

Table 4: Cyclin-dependent kinase 2 active and decoy ligand docking to representative structures of CDK: actives recovered, enrichment factors, and chemotype count for each structure at the 1% and 4% level

	Count of active ligands		Enrichment factor		Chemotype count	
Cluster representative	Top 1%	Top 4%	EF (1%)	EF (4%)	Top 1%	Top 4%
1OI9	10 (10)	18	20.0	9.0	6	11
2BHE	7 (3)	11	14.0	5.5	5	6
1GZ8	4 (4)	11	8.0	5.5	3	5
1W0X	3 (1)	10	6.0	5.0	2	6
Ensemble		18		9.0		11

Refer to Table 2 for a description of the data.

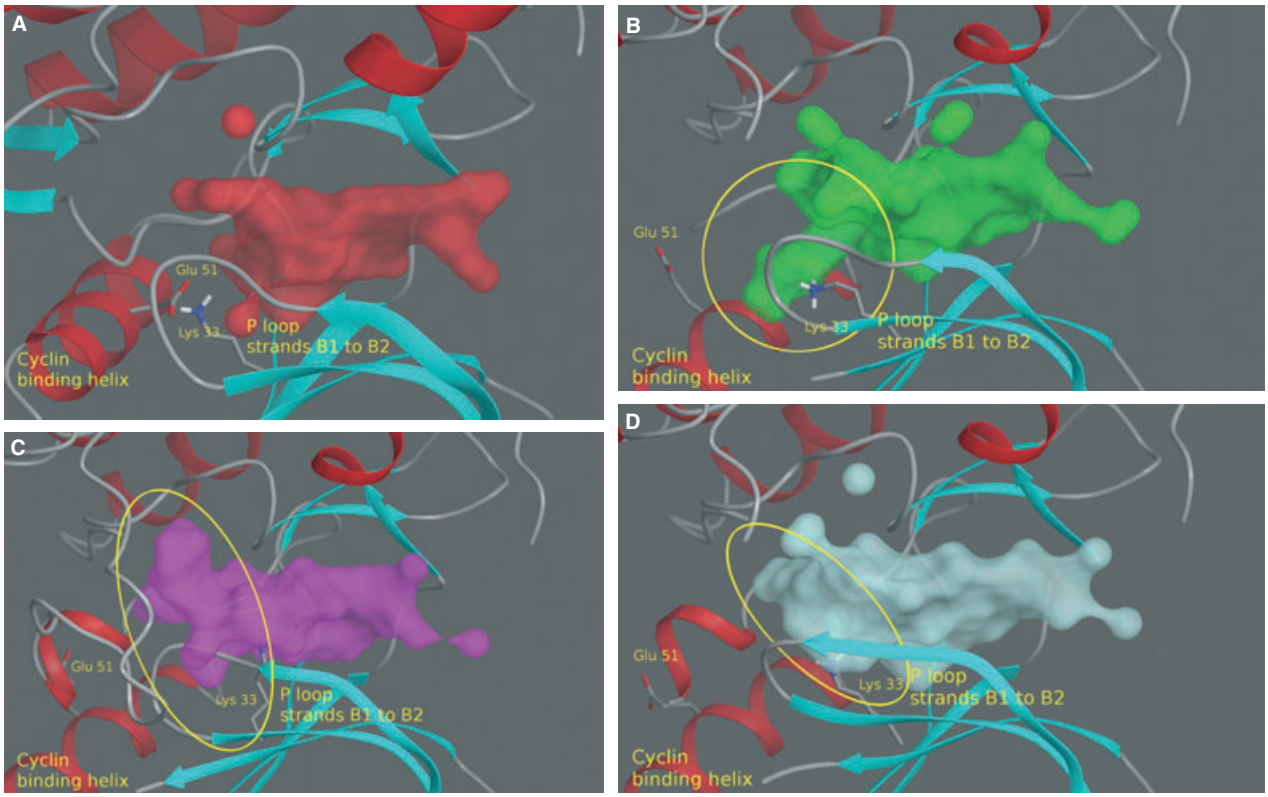

Figure 5: Ligand binding site shapes for the four cluster representatives of cyclin-dependent kinase 2 (CDK2), PDB IDs 1OI9, 2BHE, 1GZ8, and 1W0X, respectively. The yellow circle and ellipses show where the main shape change occurs compared to structure (A), which is smaller than the other volumes and mostly a subset of these. The side chains are shown for two residues, Gln 51 and Lys 33 (in the lower left quadrant), which interact with phosphates of the ATP moiety that normally binds in this site.

average of 12.5 hits obtained in the highest ranking 4% of single structures). In this case, the top ranking 4% of the compounds resulting from the docking into one CDK2 structure (Cluster representative 1) also recovers 18 actives. In addition, the diversity analysis shows that the shape-based binding site diversity method yields as many chemotypes as from any single structure and significantly more than the average single-target results. Unfortunately, as discussed in the HIV protease results, it is challenging to know *a priori* which single structure might perform best in virtual screening, thereby re-emphasizing the value of the shape-based binding site clustering approach presented here.

Androgen receptor molecular dynamics analysis

To determine the effectiveness of exploiting MD to generate the small ensemble of structures for docking, we made use of a trajectory from an MD study of the AR (53). Structures were selected at 5 ps intervals from 4 to 5 ns of the MD simulation of an AR–DHT complex, resulting in a sample of 200 structures. These structures were then clustered in the same way as the HIV protease and CDK2 crystal structures discussed above. Figure 6 shows the four representative structures from the MD along with the starting crystal structure. The clustering dendrogram is given in Figure S3.

The MD structures (B–E in Figure 6) show distinctly different shapes from the original crystal structure (Figure 6A). The MD structures cluster into effectively two different shapes, which can also be seen in Table 5 where the normalized overlap volumes between each of these pairs (1–2; 0.71 and 3–4; 0.69) are greater than the normalized overlap volumes between any structure of the two pairs, which range from 0.55 to 0.66. The results are interesting in that the structural flexibility sampled by the MD here is less than that sampled by crystal structures of diverse ligands as reflected in the overlap volumes (>0.5 in AR as opposed to approximately 0.4 for the crystal systems). This likely indicates that the sampling of configurational space is incomplete in the 1 ns sampled from the trajectory. On the contrary, it may also be that AR is less flexible than HIV or CDK2. We suspect the former, and the determination of sampling adequacy is the subject of ongoing research.

Androgen receptor virtual screening

The results of docking known active ligands and DUD decoys for the AR are given in Table 6, which shows data for docking into the single crystal structure (PDB ID 1T63) and the four cluster representative structures from the MD trajectory (of the same crystal structure). Interestingly, better results are obtained (i.e., higher EFs and more diversity) for each of the four MD structures than for the crystal structure. This could indicate that the crystal structure is highly adapted to the crystal ligand and, therefore, cannot accommodate the full diversity of active ligands.

Table 6 also shows the results of combining the top (1% of the) hits from each of the four representative structures with those obtained were we to dock into a single MD structure. Here, we see that a dramatically improved enrichment results from using a set of diverse structures (in this case from MD) as opposed to a single crystal structure, as in the CDK2 and HIV systems. However, in the MD case, unlike the crystal systems discussed above, the number of unique active ligands and EFs obtained from combining the top 1% from each of the four structures are not quite as good as the average 4% results. Nonetheless, as was observed in the case of CDK2, one of the MD structures (cluster representative 3) yields significantly more hits than the other structures. Also as noted in the discussion of the CDK2 results, the selection of an individual optimal structure is a significant challenge to the field.

We note that even if we exclude the third structure, the number of hits achieved by docking into the individual MD structures is still comparable to the number of hits retrieved from the combined set. Thus, in this case, significant improvement of enrichment is not achieved using the binding site shape-based diversity analysis. This may be correlated with the similarities of the binding site shapes, as reflected in the overlap matrices. However, despite this, of equal or perhaps more importance than the number of hits is the robustness of the results and diversity of hits. One of the key goals of virtual screening studies is to obtain new chemotypes reliably. Here, the combined results yield a larger number and more diverse set of hits than the average 'quality' of the hits from docking into the individual structures, as measured by the diversity of the sets. In fact, both the crystal and two of the MD structures only find one 'chemotype' at 1%, and the top actives from the crystal and one of the MD structures only contain this one 'chemotype' even at 4%. Thus, a better quality of hits is achieved from combining the results from diverse structures. The low level of diversity represented in the AR hits is distinct from the HIV and CDK2 structures where the diversity recovered is around 50% of the total diversity in all actives, compared to only 33% for the ensemble result for AR.

Conclusions

We have presented a procedure to select a small conformational ensemble of protein structures based on clustering of binding site volume overlaps to find diverse binding site shapes that can be used in virtual screening. The hypothesis was that an appropriate ensemble of structures would improve enrichment and diversity of retrieved active compounds. The procedure developed in this work is systematic and does not require subjective decisions about which structures to use, offering a significant step forward in the inclusion of receptor flexibility in virtual screening. In previous works, structural ensembles have been exhaustively analyzed to see if structures exist that can produce good enrichments (47), but an objective criterion to extract a useful ensemble of structures has only been validated on a single target (61) or with prior knowledge about the ligands that bind to different structures in the ensemble (62).

The HIV and CDK2 docking results using crystal structures showed that, on average, database enrichment and ligand diversity are improved using the procedure described here. The results for AR showed that if the availability of experimental crystal structures is limited, then MD can significantly improve compound selection and ligand diversity compared with the crystal structure. However, the same level of enhanced enrichment was not achieved using an MD ensemble over a single MD structure. One possible explanation for

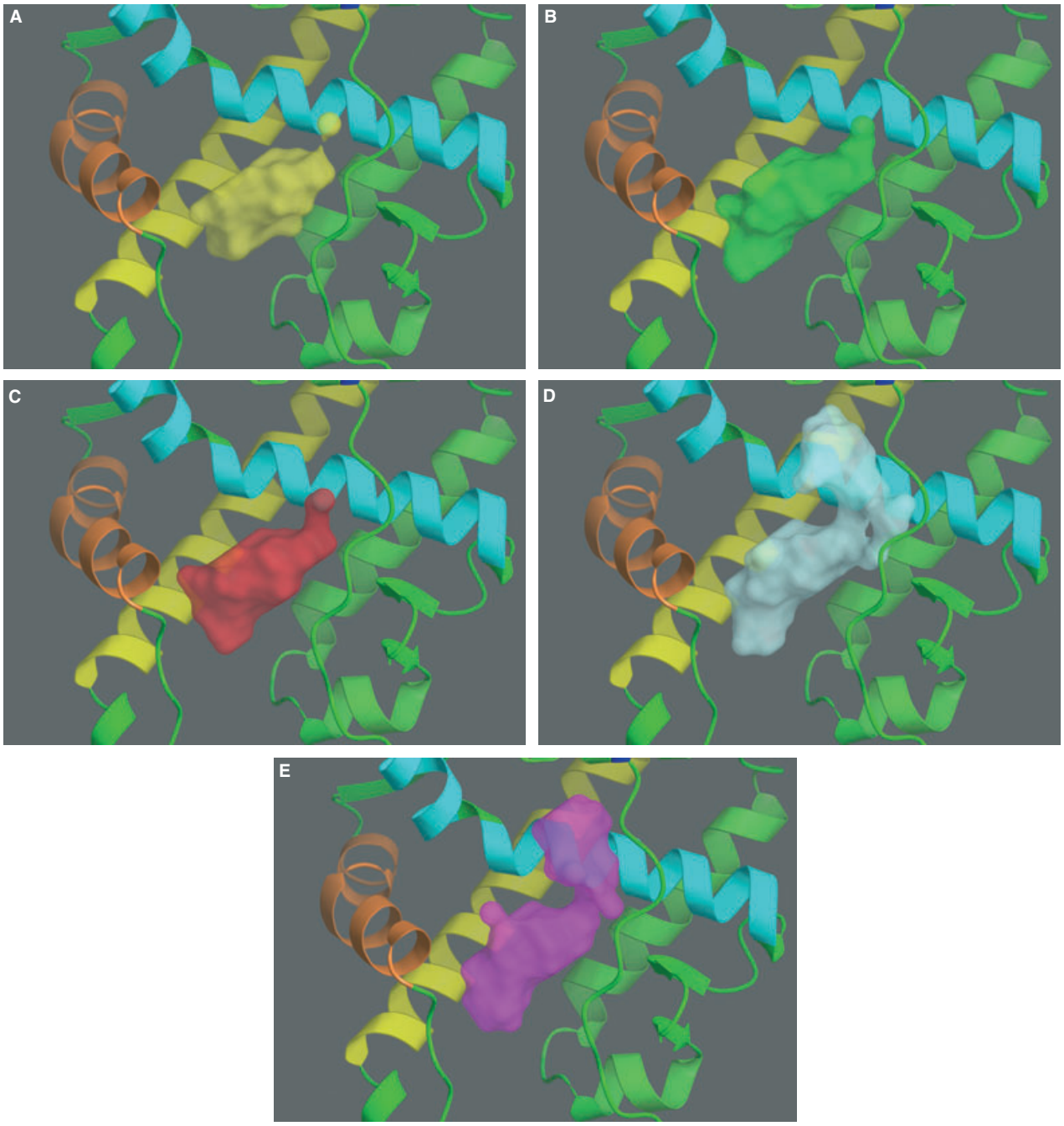

Figure 6: Active site shapes for the crystal structure of androgen receptor (AR) (top) and the four cluster representatives from molecular dynamics (MD). Helix 3 of AR has been removed from view for clarity. The yellow surface (A) is the active site shape for the 1T63 crystal structure. The green, red, cyan, and magenta surfaces (B,C,D,E) are the active site shapes for cluster representatives 1, 2, 3, and 4, respectively.

this is that the MD ensembles do not sample the range of states observed in crystal structures. This hypothesis is tested in a related publication from our group by running much longer simulations using replica exchange molecular dynamics (REMD) (63). Nevertheless, the diversity of the hits recovered from the ensemble docking work here was superior to that from a single MD structure. Overall, the binding site shape diversity procedure to generate structural ensembles offers a relatively straightforward and efficient way to include protein flexibility into the structure-based virtual screening process.

Although the results presented here are encouraging, the dataset is limited and more examples will be needed to show the robustness and applicability across different target classes where varying

Table 5: Normalized overlap volume matrix of the four molecular dynamics cluster representatives for androgen receptor

	1	2	3	4
1	1.00			
2	0.71	1.00		
3	0.59	0.55	1.00	
4	0.66	0.61	0.69	1.00

Table 6: Androgen receptor active ligands and decoy ligands docking to a crystal structure and four molecular dynamics representative structures

Structure or cluster representative	Count of active ligands		Enrichment factor		Chemotype count	
	Top 1%	Top 4%	EF (1%)	EF (4%)	Top 1%	Top 4%
Crystal (1T63)	12	23	16.2	7.8	1	1
1	18 (18)	25	24.3	8.4	1	2
2	20 (4)	26	27.0	8.8	1	1
3	20 (3)	32	27.0	10.8	3	4
4	15 (0)	25	20.3	8.4	2	2
Ensemble		25		8.4		3

Active ligand counts, enrichment factors, and chemotype count for each structure at the 1% and 4% level. See Table 2 for a description of the data.

degrees of protein flexibility exist. Furthermore, there has been no attempt in this work (or in other virtual screening papers, to our knowledge) to account for the protein strain energy associated with each structure. The direct calculation of protein strain energy using physics-based methods is an immensely challenging problem for the field of computational chemistry and is beyond the scope of this work. However, it may be possible to parameterize protein strain energies using experimental data, either directly from binding affinity measurements or from structural populations observed in solution spectroscopy, such as NMR, when available. This remains an important area of research for the field. Within the framework presented here, each protein structure is treated as equienergetic, which appears to be an acceptable approximation even if more accurate strain energy predictions could produce better results.

Future work will involve more sophisticated methods to choose protein conformations. Although the shape diversity approach presented here was shown to work well, additional descriptors could be used to improve the results. For example, it is possible that a snapshot from a MD simulation has a collapsed binding site that precludes active ligands from binding. In such a case, the shape-based diversity selection could choose this collapsed structure as being different from the other structures, although this is unlikely to be a useful structure to include in the structural ensemble for docking. We are currently looking at complementing the diversity with the druggability score (DScore) from SiteMap to focus on binding site conformations that are both diverse and highly druggable. The results of this study will be presented in future work. In addition, the MD simulations performed here were relatively short and no attempt was made to optimize the MD protocol. In another publication (63) we showed that much longer simulations with REMD to enhance sampling resulted in only a modest improvement in enrichments, suggesting that the protocol presented here is a good starting point for including a small amount of receptor flexibility in the structure-based virtual screening process. In summary, the work described here represents a step forward toward including protein flexibility in virtual screening, but there is still a significant amount of work to do to create an optimum solution that reliably produces an improvement in enrichment and diversity of hits across a broad range of targets.

References

1. Shoichet B.K. (2004) Virtual screening of chemical libraries. Nature;432:862–865.
2. Walters W.P., Stahl M.T., Murcko M.A. (1998) Virtual screening – an overview. Drug Discov Today;3:160–178.
3. Stahl M., Rarey M. (2001) Detailed analysis of scoring functions for virtual screening. J Med Chem;44:1035–1042.
4. Noha S.M., Atanasov A.G., Schuster D., Markt P., Fakhrudin N., Heiss E.H., Schrammel O., Rollinger J.M., Stuppner H., Dirsch V.M., Wolber G. (2011) Discovery of a novel ikk-β inhibitor by ligand-based virtual screening techniques. Bioorg Med Chem Lett;21:577–583.
5. Miteva M.A., Lee W.H., Montes M.O., Villoutreix B.O. (2005) Fast structure-based virtual ligand screening combining fred, dock, and surflex. J Med Chem;48:6012–6022.
6. Muegge I., Enyedy I.J. (2004) Virtual screening for kinase targets. Curr Med Chem;11:693–707.
7. Bissantz C., Logean A., Rognan D. (2004) High-throughput modeling of human g-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening. J Chem Inf Comput Sci;44:1162–1176.
8. Duan J., Dixon S.L., Lowrie J.F., Sherman W. (2010) Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. J Mol Graph Model;29:157–170.
9. Sastry M., Lowrie J.F., Dixon S.L., Sherman W. (2010) Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. J Chem Inf Model;50:771–784.
10. Singh J., Chuaqui C.E., Boriack-Sjodin P., Lee W.C., Pontz T., Corbley M.J., Cheung H.K., Arduini R.M., Mead J.N., Newman M.N. (2003) Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I tgfβ receptor kinase (tβri). Bioorg Med Chem Lett;13:4355–4359.
11. Lyne P.D. (2002) Structure-based virtual screening: an overview. Drug Discov Today;7:1047–1055.
12. Bissantz C., Folkers G., Rognan D. (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem;43:4759–4767.
13. Salam N.K., Huang T.H., Kota B.P., Kim M.S., Li Y., Hibbs D.E. (2008) Novel ppar-gamma agonists identified from a natural product library: a virtual screening, induced-fit docking and biological assay study. Chem Biol Drug Des;71:57–70.

14. Good A.C., Cheney D.L., Sitkoff D.F., Tokarski J.S., Stouch T.R., Bassolino D.A., Krystek S.R., Li Y., Mason J.S., Perkins T.D. (2003) Analysis and optimization of structure-based virtual screening protocols. 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success. J Mol Graph Model;22:31–40.
15. Pierce A.C., Jacobs M., Stuver-Moody C. (2008) Docking study yields four novel inhibitors of the protooncogene pim-1 kinase. J Med Chem;51:1972–1975.
16. Foloppe N., Fisher L.M., Howes R., Potter A., Robertson A.G.S., Surgenor A.E. (2006) Identification of chemically diverse chk1 inhibitors by receptor-based virtual screening. Bioorg Med Chem;14:4792–4802.
17. 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. (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem;47:1739–1749.
18. Kellenberger E., Rodrigo J., Muller P., Rognan D. (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins;57:225–242.
19. Kroemer R.T., Vulpetti A., McDonald J.J., Rohrer D.C., Trosset J.Y., Giordanetto F., Cotesta S., McMartin C., KihlÈn M., Stouten P.F.W. (2004) Assessment of docking poses: interactions-based accuracy classification (ibac) versus crystal structure deviations. J Chem Inf Comput Sci;44:871–881.
20. Thomsen R., Christensen M.H. (2006) Moldock: a new technique for high-accuracy molecular docking. J Med Chem;49:3315–3321.
21. Jain A.N. (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem;46:499–511.
22. 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. (2006) A critical assessment of docking programs and scoring functions. J Med Chem;49:5912–5931.
23. Halgren T.A., Murphy R.B., Friesner R.A., Beard H.S., Frye L.L., Pollard W.T., Banks J.L. (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem;47:1750–1759.
24. Zsoldos Z., Reid D., Simon A., Sadjad S.B., Johnson A.P. (2007) Ehits: a new fast, exhaustive flexible ligand docking system. J Mol Graph Model;26:198–212.
25. Sherman W., Day T., Jacobson M.P., Friesner R.A., Farid R. (2006) Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem;49:534–553.
26. Nabuurs S.B., Wagener M., de Vlieg J. (2007) A flexible approach to induced fit docking. J Med Chem;50:6507–6518.
27. Corbeil C.R., Englebienne P., Moitessier N. (2007) Docking ligands into flexible and solvated macromolecules. 1. Development and validation of fitted 1.0. J Chem Inf Model;47:435–449.
28. Rueda M., Bottegoni G., Abagyan R. (2009) Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. J Chem Inf Model;49:716–725.
29. Park S.J., Kufareva I., Abagyan R. (2010) Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. J Comput Aided Mol Des;24:459–471.
30. Cavasotto C.N., Abagyan R.A. (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. J Mol Biol;337:209–225.
31. Sherman W., Beard H.S., Farid R. (2006) Use of an induced fit receptor structure in virtual screening. Chem Biol Drug Des;67:83–84.
32. Rueda M., Bottegoni G., Abagyan R. (2009) Recipes for the selection of experimental protein conformations for virtual screening. J Chem Inf Model;50:186–193.
33. Brooks B.R., Bruccoleri R.E., Olafson B.D., States D.J., Swaminathan S., Karplus M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem;4:187–217.
34. Pearlman D.A., Case D.A., Caldwell J.W., Ross W.S., Cheatham T.E. (1995) Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput Phys Commun;91:1–41.
35. Phillips J., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R., KalÈ L., Schulten K. (2005) Scalable molecular dynamics with namd. J Comb Chem;26:1781.
36. Bowers K.J., Chow E., Xu H., Dror R.O., Eastwood M.P., Gregersen B.A., Klepeis J.L., Kolossváry I., Moraes M.A., Sacerdoti F.D., Salmon J.K., Shan Y., Shaw D.E. (2006) Scalable algorithms for molecular dynamics simulations on commodity clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06). Tampa, Florida, November 11–17. The program authors request on their webpage that this work is referenced even though it is a conference proceeding: http://www.deshawresearch.com/resources_desmondpublications.html.
37. Berendsen H., van der Spoel D., Van Drunen R. (1995) Gromacs: a message-passing parallel molecular dynamics implementation. Comput Phys Commun;91:43–56.
38. Sugita Y., Okamoto Y. (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett;314:141–151.
39. Karplus M., Petsko G.A. (1990) Molecular dynamics simulations in biology. Nature;347:631–639.
40. Freddolino P.L., Liu F., Gruebele M., Schulten K. (2008) Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. Biophys J;94:L75–L77.
41. Klepeis J.L., Lindorff-Larsen K., Dror R.O., Shaw D.E. (2009) Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol;19:120–127.
42. Klein M.L., Shinoda W. (2008) Large-scale molecular dynamics simulations of self-assembling systems. Science;321:798.
43. McCammon J.A., Gelin B.R., Karplus M. (1977) Dynamics of folded proteins. Nature (London);267:585–590.
44. Cheng L.S., Amaro R.E., Xu D., Li W.W., Arzberger P.W., McCammon J.A. (2008) Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. J Med Chem;51:3878–3894.
45. Ferrara P., Curioni A., Vangrevelinghe E., Meyer T., Mordasini T., Andreoni W., Acklin P., Jacoby E. (2005) New scoring functions for virtual screening from molecular dynamics simulations with a quantum-refined force-field (qrff-md). Application to cyclin-dependent kinase 2. J Chem Inf Model;46:254–263.

46. Kollman P.A., Massova I., Reyes C., Kuhn B., Huo S., Chong L., Lee M., Lee T., Duan Y., Wang W., Donini O., Cieplak P., Srinivasan J., Case D.A., Cheatham T.E. (2000) Calculating structures and free energies of complex molecules:'äâ combining molecular mechanics and continuum models. Acc Chem Res;33:889–897.
47. Nichols S.E., Baron R., Ivetac A., McCammon J.A. (2011) Predictive power of molecular dynamics receptor structures in virtual screening. J Chem Inf Model;51:1439–1446.
48. Totrov M., Abagyan R. (2008) Flexible ligand docking to multiple receptor conformations: a practical alternative. Curr Opin Struct Biol;18:178–184.
49. Kitchen D.B., Decornez H., Furr J.R., Bajorath J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov;3:935–949.
50. Armen R.S., Chen J., Brooks C.L. (2009) An evaluation of explicit receptor flexibility in molecular docking using molecular dynamics and torsion angle molecular dynamics. J Chem Theory Comput;5:2909–2923.
51. Halgren T. (2009) Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model;49:377–389.
52. Halgren T. (2007) New method for fast and accurate binding-site identification and analysis. Chem Biol Drug Des;69:146–148.
53. Osguthorpe D.J., Hagler A.T. (2011) Mechanism of androgen receptor antagonism by bicalutamide in the treatment of prostate cancer. Biochemistry;50:4105–4113. DOI: 10.1021/bi102059z
54. Estebanez-Perpina E., Moore J.M., Mar E., Delgado-Rodrigues E., Nguyen P., Baxter J.D., Buehrer B.M., Webb P., Fletterick R.J., Guy R.K. (2005) The molecular mechanisms of coactivator utilization in ligand-dependent transactivation by the androgen receptor. J Biol Chem;280:8060–8068.
55. Eldridge M.D., Murray C.W., Auton T.R., Paolini G.V., Mee R.P. (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des;11:425–445.
56. Durant J.L., Leland B.A., Henry D.R., Nourse J.G. (2002) Reoptimization of mdl keys for use in drug discovery. J Chem Inf Comput Sci;42:1273–1280.
57. Salam N.K., Nuti R., Sherman W. (2009) Novel method for generating structure-based pharmacophores using energetic analysis. J Chem Inf Model;49:2356–2368.
58. Huang N., Shoichet B.K., Irwin J.J. (2006) Benchmarking sets for molecular docking. J Med Chem;49:6789–6801.
59. Specker E., Bottcher J., Lilie H., Heine A., Schoop A., Muller G., Griebenow N., Klebe G. (2005) An old target revisited: two new privileged skeletons and an unexpected binding mode for hiv-protease inhibitors. Angew Chem Int Ed Engl;44: 3140–3144.
60. Rutenber E.E., McPhee F., Kaplan A.P., Gallion S.L., Hogan J.C. Jr, Craik C.S., Stroud R.M. (1996) A new class of hiv-1 protease inhibitor: the crystallographic structure, inhibition and chemical synthesis of an aminimide peptide isostere. Bioorg Med Chem;4:1545–1558.
61. Rao S., Sanschagrin P.C., Greenwood J.R., Repasky M.P., Sherman W., Farid R. (2008) Improving database enrichment through ensemble docking. J Comput Aided Mol Des;22:621–627.
62. Ferrari A.M., Wei B.Q., Costantino L., Shoichet B.K. (2004) Soft docking and multiple receptor conformations in virtual screening. J Med Chem;47:5076–5084.
63. Osguthorpe D.J., Sherman W., Hagler A.T. (2012) Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. J Phys Chem B;???:???–???. Article ASAP DOI: 10.1021/jp3003992.

Notes

[a]Abola E.E., Bernstein F.C., Bryant S.H., Koetzle T.F., Weng J. (1987) Protein data bank. Crystallograpic Databases – Information Content, Software Systems, Scientific Applications. Sect. 107–132.
[b]Tanimoto T. IBM internal report 1957. November; 1957.
[c]Schrodinger L.L.C. Canvas. 1.4 ed. New York, NY2011.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Dendrogram of ligand binding site shape clustering for 135 HIV protease structures.

Figure S2. Dendrogram of ligand binding site shape clustering from 92 CDK2 structures.

Figure S3. Dendrogram of ligand binding site shape clustering from 200 AR structures from 1 ns of molecular dynamics sampled every 5 ps.

Figure S4. Histograms of all pairwise volume overlaps from the clustering matrices for HIV protease (top), CDK2 (middle), and AR (bottom).

Figure S5. ROC curves for each individual structure within each ensemble for the three targets.

Figure S6. ROC curves for each of the three structural ensembles.

Appendix S1. PDB IDs of all structures.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.