

Article

## Exploring Protein Flexibility: Incorporating Structural Ensembles From Crystal Structures and Simulation into Virtual Screening Protocols

David J. Osguthorpe, Woody Sherman, and Arnie T. Hagler

*J. Phys. Chem. B*, **Just Accepted Manuscript** • DOI: 10.1021/jp3003992 • Publication Date (Web): 16 Mar 2012

Downloaded from <http://pubs.acs.org> on April 15, 2012

### Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.



ACS Publications  
High quality. High impact.

The Journal of Physical Chemistry B is published by the American Chemical Society.  
1155 Sixteenth Street N.W., Washington, DC 20036  
Published by American Chemical Society. Copyright © American Chemical Society.  
However, no copyright claim is made to original U.S. Government works, or works  
produced by employees of any Commonwealth realm Crown government in the course  
of their duties.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# Exploring Protein Flexibility: Incorporating Structural Ensembles From Crystal Structures and Simulation into Virtual Screening Protocols

*David J. Osguthorpe,<sup>1</sup> Woody Sherman,<sup>2</sup> and Arnold T. Hagler<sup>1, 3\*</sup>*

1. Shifa Biomedical, 1 Great Valley Parkway, Suite 8, Malvern, Pennsylvania 19355, United States

2. Schrödinger, Inc., 120 West 45<sup>th</sup> Street, 17<sup>th</sup> Floor, New York, NY 10036

3. Department of Chemistry, University of Massachusetts, 701 Lederle Graduate Research Tower, 710 North Pleasant Street, Amherst, Massachusetts 01003-9336, United States

\*A.T. Hagler, Dept. of Chemistry, University of Massachusetts, PO Box 12067 La Jolla Ca 92039,  
Phone: 619 379-9768, e-mail: athagler@gmail.com

## Abstract

The capacity of proteins to adapt their structure in response to various perturbations including covalent modifications, and interactions with ligands and other proteins plays a key role in biological processes. Here, we explore the ability of molecular dynamics (MD), replica exchange molecular dynamics (REMD) and a library of structures of crystal-ligand complexes, to sample the protein conformational landscape and especially the accessible ligand binding site geometry. The extent of conformational space sampled is measured by the diversity of the shapes of the ligand binding sites. Since our focus here is the effect of this plasticity on the ability to identify active compounds through virtual screening, we use the structures generated by these techniques to generate a small ensemble for further docking studies, using binding site shape hierarchical clustering to determine four structures for each ensemble. These are then assessed for their capacity to optimize enrichment and diversity in docking. We test these protocols on three different receptors: androgen receptor (AR), HIV protease, and CDK2. We show that REMD enhances structural sampling slightly as compared both to MD, and the distortions induced by ligand binding as reflected in the crystal structures. The improved sampling of the simulation methods doesn't translate directly into improved docking performance however. The ensemble approach did improve enrichment and diversity, and, the ensemble derived from the crystal structures performed somewhat better than those derived from the simulations.

**Keywords:** Molecular dynamics, REMD, clustering, enrichment, diversity, binding site

## Introduction

In a previous paper we investigated the use of a small ensemble of protein structures based on the diversity of binding site shapes to account for protein flexibility in docking. We found that the use of four diverse structures indeed improves consistency in results, database enrichment rates, and diversity of the hits obtained from docking. The studies were performed on crystal structures of HIV-1 protease and CDK2 and also on a structural ensemble generated from molecular dynamics (MD) simulations of androgen receptor (AR). We found that the MD ensemble of structures produced improved docking enrichment over a single crystal structure, consistent with the findings from the crystal structure ensembles. However, using multiple MD structures for AR did not improve results as significantly as using multiple crystal structures did for the HIV-1 protease and CDK2 systems and, perhaps surprisingly, the diversity of shapes was smaller than from the different crystal structures.

In this paper we apply MD to the same three systems (HIV-1 protease, CDK2, and AR) and study more rigorous sampling through the use of replica exchange molecular dynamics (REMD)<sup>1</sup>. The diversity of ligand binding site (LBS) shapes and the enrichment achieved from docking into the ensemble of structures obtained in these studies is compared to those obtained from the previous crystal studies. In addition we generated an ensemble of ligand binding shapes from the universe of AR-ligand crystal structures and carried out the same analysis of docking results using these.

Our objective is twofold. First, we explore whether the use of MD to improve enrichment through generating diverse structures is a generally applicable approach - or if the AR results were in some way unique to that system. Secondly, we investigate whether the apparently limited sampling of ligand binding site structure by the MD was due to the use of a limited portion of a single trajectory or is also a general phenomenon. REMD was performed to improve the sampling over standard MD with the hopes that the additional sampling from the single initial crystal structure would improve the virtual screening

1 results. In addition, we performed an LBS SiteMap clustering of the REMD “trajectories” as well as on  
2  
3 44 AR crystal structures to have a complete set of the three systems. The results of docking into the  
4  
5 small ensemble of diverse LBS structures resulting from this clustering was then analyzed to ascertain if  
6  
7 the enrichment and diversity achieved with REMD is comparable or perhaps superior to that achieved  
8  
9 from crystal structures or MD.  
10

## 11 12 13 14 Methods

15  
16  
17  
18  
19 *Crystal Structures and Initial Preparation.* We searched the PDB for AR ligand binding domain  
20  
21 (LBD) structures and selected 44 for clustering. The protein structures were prepared using the  
22  
23 Schrodinger Protein Preparation Wizard (PrepWizard) <sup>2</sup>. This preparation protocol added hydrogens,  
24  
25 built side chains and loops with missing atoms, determined the optimal protonation states for ionizable  
26  
27 residues, optimized the hydrogen-bonding network, and performed a restrained minimization to obtain  
28  
29 the final structure for running docking or simulations.  
30  
31  
32  
33  
34

35  
36 *Choice of Systems for MD Studies.* As noted in the introduction, we studied the three systems (HIV-1  
37  
38 protease, CDK2, and AR) used previously to assess database enrichments and diversity of retrieved  
39  
40 ligands in ensemble docking. For HIV-1 protease, we used the 1EBZ structure from the PDB for the MD  
41  
42 since 1EBZ was the representative structure of the cluster with the largest number of structures from our  
43  
44 previous clustering of 135 HIV-1 protease structures <sup>3</sup>. Two CDK2 structures were chosen, one with  
45  
46 cyclin bound and another without cyclin bound. These two forms have significant structural changes due  
47  
48 to re-folding of the sections of the CDK2 structure involved in cyclin binding. Our previous clustering  
49  
50 of 92 structures <sup>3</sup> separated these two forms into different clusters, and we chose the non-cyclin bound  
51  
52 structure as the cluster representative from the largest cluster of non-cyclin bound structures (1W0X)  
53  
54 and the cluster representative of the single cyclin bound cluster (1OI9) for the MD simulations. For AR  
55  
56 we used the 1T63 AR-DHT complex we investigated previously <sup>4</sup>.  
57  
58  
59  
60

*Preparation of Protein Systems for MD and REMD.* The PDB structures were prepared with the Schrodinger PrepWizard utility as described above. The prepared systems were solvated using the System Builder program of the Desmond suite with a dodecahedral solvent box and a solvent buffer extending 10 Å beyond the protein in all directions. The systems were neutralized with counterions, which entailed adding different numbers of Na<sup>+</sup> or Cl<sup>-</sup> ions to each system. 4 Cl<sup>-</sup> ions were added for HIV-1 protease, 5 and 2 Cl<sup>-</sup> ions were added for apo and cyclin bound CDK2 respectively. For the AR structure, we used the existing prepared starting structure for the solvated MD studies as in our previously published AR work <sup>4</sup>.

*Molecular Dynamics.* MD simulations were carried out using the Desmond suite <sup>5</sup> with long-range electrostatic interactions computed using a smooth Particle-Mesh Ewald (PME) approximation <sup>6</sup> with a cutoff radius of 9.0 Å for the transition between the particle-particle and particle-grid calculations. Van der Waals (vdW) interactions were truncated at 9.0 Å. Dodecahedral periodic boundary conditions were used for all simulations. MD steps were integrated using a two time-step algorithm, with 2 fs steps for bonded and short-range interactions within the 9.0 Å cutoff and 10 fs for long-range non-bonded interactions. The system was relaxed using a protocol consisting of an initial minimization, MD for 20 ps at 0.1K, MD for 20 ps at 310K both with restraints of 250 kcal/mol/Å<sup>2</sup> on all heavy atoms using the Berendsen <sup>7</sup> thermostat followed by 100 ps with restraints of 2.5 kcal/mol/Å<sup>2</sup> performed using 1 fs short range and 5 fs long range timesteps to reduce numerical issues with large initial forces. This was followed by 100 ps using the NPT ensemble and Berendsen thermostat and barostat and 100 ps Nose-Hoover thermostat and Martyna-Tuckerman-Klein barostat <sup>8</sup>. This was extended to 1ns with no restraints. The production MD simulations were carried out at 310 K using the NVT ensemble with a Nose-Hoover thermostat <sup>9,10</sup>. Trajectories were run for 5 ns, with the equilibration being carried out over the first nanosecond and analysis carried out on the final four nanoseconds.

*Replica Exchange Molecular Dynamics.* REMD is a procedure for improving sampling by performing multiple simulations of the same system at different temperatures in parallel<sup>1,11-13</sup>. At fixed time intervals a so-called “exchange” is attempted in which both the coordinates and velocities from a pair of adjacent temperature simulations are swapped. A Metropolis criterion<sup>14</sup> is used to determine if the exchange should be accepted or not (eqn 1).

$$P(x | x') = \begin{cases} 1 & \text{for } \Delta \leq 0 \\ \exp(-\Delta) & \text{for } \Delta > 0 \end{cases}$$

where  $\Delta = (\beta - \beta')(V(x') - V(x))$

$$\beta = 1/kT, \quad \beta' = 1/kT' \quad (1)$$

$x$  and  $x'$  are the configurations at  $T$  and  $T'$  respectively

$V$  and  $V'$  are the potential energies of the two configurations

This procedure guarantees that in the long run for each simulation in the REMD set, averages over the trajectory are thermodynamically valid. The ability for configurations from higher temperatures to swap to lower temperatures improves the sampling of configurations at that temperature compared to performing a standard molecular dynamics trajectory.

The rate at which exchanges are accepted should be similar for each trajectory and allow a reasonable residence time for an exchanged structure at that temperature (around 10 ns). The temperature differences needed to achieve this can be estimated using an algorithm published by Patriksson<sup>15</sup>. This algorithm was used to determine the temperature distribution in this work. Given our initial choice of a 100-degree temperature range, from 298-398 K, the total number of replicas was determined as a function of the system size (i.e. number of atoms). The 100-degree temperature range was intended to improve sampling but not to introduce unfolding at the higher temperatures. Forty-five replicas were required for HIV-1 protease, while 55 and 53 replicas were used for the two CDK2 systems and AR required 44 replicas. REMD simulations at each temperature were performed for 3 ns using the

Desmond program suite<sup>5</sup> with the same methods as described above for the MD, resulting in approximately 150 ns of total simulation time for each system.

*Ligand Binding Site Shape Clustering.* As in our previous study<sup>3</sup>, the program SiteMap<sup>16,17</sup> was used to compute the binding site shape for each receptor structure. Structures were superimposed using C $\alpha$  atoms of common residues within 8 Å of the ligands. The volume overlap of the sites for each pair of structures was computed to create the volume overlap matrix. Hierarchical clustering was applied to this matrix using average linkage to create a clustering order. A cluster partitioning of four was chosen as a balance between time and completeness of sampling. For each partition the cluster representative was determined to realize a set of diverse binding site shapes.

*Ligand Datasets.* The DUD dataset<sup>18</sup> was chosen as the test dataset to provide active and decoy ligands. This included 52 unique active ligands and 1884 decoys for HIV-1 protease and 49 unique active ligands and 1778 decoys for CDK2. For androgen receptor (AR), the set included 73 active ligands and 2627 decoys.

*Docking.* As in our previous work<sup>3</sup>, the Glide SP algorithm was used for docking<sup>19,20</sup> with the final scoring using the GlideScore. The GlideScore has empirical terms to account for desolvation effects and special reward terms to account for interactions that are difficult to accurately compute with force fields, such as pi-pi and pi-cation interactions.

*Database Enrichment and Diversity Assessment.* The enrichment factor (EF) was computed using eqn 2:

$$EF = F_d/F_d \quad (2)$$



where  $F_a$  is the fraction of actives found and  $F_d$  is the fraction of the database sampled. In this study, as previously, we have calculated enrichment factors for the actives found in the top scoring 1% and 4% of the total compounds docked (denoted EF (1%) and EF (4%), respectively).

The diversity of ligands selected was computed as follows: The MACCs fingerprints<sup>21</sup> of the active ligands were determined and hierarchical clustering was performed on the matrix of pair similarities determined using the Tanimoto metric using the program Canvas<sup>22</sup>. From inspection of the resulting clustering of all actives in each DUD ligand dataset (figure 1 shows the clustering of AR with those of HIV and CDK2 in supplementary material figures S4 and S5), a clustering level of 0.7 similarity was chosen as separating the active ligands into different chemotypes. For each cluster representative, after docking the actives in the 1% ligands were clustered and the number of clusters at the 0.7 level was determined. This number of clusters gives an estimate of the diversity of the hits found from docking.

## Results

*Molecular dynamics.* In order to assess the ability of molecular dynamics (MD) to sample the accessible configuration space of the ligand binding site, we carried out standard MD and replica exchange MD (REMD) simulations of AR, CDK2, and HIV-1 protease. The MD and REMD trajectories of the distance between residues roughly characterizing the LBS width are given in Figure 2 for the three proteins. As can be seen from the figure the distances visited during the trajectories for all three systems span a greater range for the REMD, as might be expected from its superior sampling ability<sup>1,11-13</sup> even though the REMD trajectories are slightly shorter. This is especially true for the HIV-1 protease and CDK2 systems, while in the case of AR the MD eventually visits almost the whole span of distances, albeit briefly in some cases.

*Exploration of LBS Shapes By MD Versus Crystal Structures.* The structures obtained from the three sources (crystal, MD, and REMD) were then used to generate structural ensembles for virtual screening through hierarchical clustering. The resulting representative structures from MD and REMD were then compared with those obtained from the crystal structures<sup>3</sup>. A sense of the ligand binding site plasticity can be obtained by examining the LBS volumes of the cluster representatives sampled by the different methods. These are given for AR in Figure 3 and for CDK2 and HIV-1 protease in supplementary material. As can be seen in Figure 3, the crystal structures appear to show a greater diversity of volumetric shape than the MD structures, with four distinctly different shapes. The MD structures show essentially two different shapes as opposed to the greater variation in the crystal structures. This is consistent with the results in

our previous study, where the volume overlap matrices indicated the crystal complexes of HIV-1 protease and CDK2 sampled the accessible LBS volumetric space more effectively than the (limited) MD did for AR<sup>3</sup> (and was one of the motivating factors for this study). The REMD cluster representatives are distinctly different from the crystal and MD volumes but show some similarities in overall shape, with Figure 3 (k) (cyan shaded) showing similar features to (b) from the crystal and MD (h), although as seen from the overlap matrix given below, similarity of features doesn't necessarily equate to a global volume similarity.

*Overlap volumes of representative AR structures.* The greater diversity of the crystal shapes is further demonstrated by the normalized overlap volume matrix for the crystal and MD cluster representatives of AR (Table 1). The overlap matrix of the representative structures from MD confirms that the two pairs of structures are more similar to each other than to the other pair (e-f, g-h, shaded yellow and light blue) with the smallest similarity between the two pairs being 0.55.

By comparison the least similar structures from the crystals have an overlap volume of 0.45. The average normalized overlap volume for all pairs is 0.64 for the MD cluster representatives versus 0.57

1 for the crystal, showing the crystal structure volumes sample a larger binding site structural space, on  
2 average, than the MD volumes in this system. The average of the normalized overlap matrix of the four  
3 cluster representatives from the REMD sampling of AR is 0.60 which is somewhat lower than the MD  
4 average (0.64) indicating that REMD has increased the diversity of volumes slightly, but still does not  
5 achieve the sampling of LBS volume induced by different ligands (average crystal overlap volume 0.57).  
6  
7  
8  
9  
10  
11

12  
13  
14 *LBS volumetric sampling of HIV-1 protease and CDK2 yield different trends than AR.* To assess the  
15 generality of this result we compare the sampling achieved by the three methods for the HIV-1 protease  
16 and CDK2 systems as well. Figures of the representative structures arising from the clustering of the  
17 crystal and dynamic trajectories are given in the supplementary material (Figures S1 and S2). Figure S1  
18 shows the HIV-1 protease REMD shapes are more varied than the MD, with more open space in most of  
19 the representatives suggesting that the REMD has searched more of conformational space. This is  
20 confirmed by the overlap matrix for the HIV-1 protease representative structures, as seen in Table 2. In  
21 this system, unlike the behavior for AR, we see that a greater sampling efficiency is indeed achieved by  
22 the REMD method, than by either the crystal structures or the MD, consistent with the trend seen in the  
23 distance trajectory in Fig 1. Thus the average of the normalized overlap volume matrices is 0.41 and  
24 0.43, and 0.32 for the crystal, MD, and REMD sampling, respectively, reflecting a significantly more  
25 extensive sampling by the REMD method.  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44

45 *Comparison of LBS volumes obtained from different methods.* As might be expected, at least for the  
46 dynamic methods, which trace the trajectory of a single structure, the “cross” overlap matrices between  
47 volumes obtained from different methods, tend to show a greater variation in volumes than those from  
48 any of the individual techniques. Thus, for example, the overlap between the first representative  
49 structure from the crystal with the fourth REMD structure is only 0.15, lower than any of the overlaps  
50 within a set. The average overlaps for representative structures from the HIV-1 protease crystal  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1 clustering with those from MD is 0.29 and for the crystal – REMD, and MD – REMD 0.28 and 0.33  
2  
3 respectively. This trend is also observed in the AR structures.  
4  
5  
6

7 *CDK2 – Cyclin vs. non-cyclin bound.* Finally, in Table 3, we give the overlap matrices for the CDK2  
8  
9 system. The CDK2 structures separate globally into two basic classes; the cyclin and non-cyclin bound  
10  
11 structures. In the crystal the former assemble into a single cluster while the non-cyclin structures are  
12  
13 spread among the remaining three clusters with the exception of two non-cyclin structures, which appear  
14  
15 in the cyclin bound cluster (Fig S3). Similar results are found in the distribution of structures from the  
16  
17 MD. Namely, all cyclin bound structures are found in a single cluster, which interestingly contains two  
18  
19 of the 800 non-cyclin structures that are visited along the MD trajectory. The REMD results are similar  
20  
21 in that again all cyclin bound structures along the REMD “trajectory” group together in a single cluster.  
22  
23 However in this case 103 of the 800 non-cyclin structures sampled in the REMD simulation group with  
24  
25 the cyclin bound cluster. The excursion of the non-cyclin structures into the cyclin bound configuration  
26  
27 space is consistent with the enhanced ability of REMD to sample the energy landscape.  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37

38 *Simulations sample greater volume diversity than crystal structures in CDK2.* As in the case of the  
39  
40 AR and HIV-1 protease systems, another measure of the plasticity of the protein sampled by the  
41  
42 different techniques is the similarity of the LBS volumes of the representative structures. Here the  
43  
44 results differ slightly from what might be expected from the nature of the clustering and from the results  
45  
46 of the AR and HIV-1 protease systems. In this case, as can be seen from the normalized overlap volumes  
47  
48 in Table 3, both the MD and REMD simulations visit more diverse volumes than induced by the ligands  
49  
50 in the crystal structures but in this case the MD simulations also span slightly more “volume space” than  
51  
52 the REMD. This is reflected in the average overlap volumes of 0.46, 0.37 and 0.40 for the crystal, MD,  
53  
54 and REMD respectively.  
55  
56  
57  
58  
59  
60

1 *Does the ensemble of structures obtained from simulations improve enrichment over docking into a*  
2 *single structure?* The purpose of these studies is to develop methods for improving enrichment in virtual  
3 screening campaigns. Thus, the bottom line is whether docking into small ensembles obtained from  
4 simulations of dynamics improve enrichment over docking into a single structure, and secondly how this  
5 enrichment compares with that achieved with ensembles derived from crystal structures.  
6  
7  
8  
9  
10

11  
12  
13  
14 In Table 4 we show the results of a docking into the individual four cluster representatives obtained  
15 from crystal structures, MD, and REMD for the HIV-1 protease system. The enrichment achieved by  
16 exploiting the ensemble is better or equal (in the case of MD) to the average enrichment, which would  
17 be achieved by taking the same number of hits from the single structures. For example, 13 actives are  
18 found in the combined hits from the top 1% of hits from each representative structure of the crystal,  
19 while if we just took the top 4% of the hits from individual structures we would only recover ~10 actives  
20 on average. As noted previously<sup>3</sup>, there are individual structures, for example representative structure 2  
21 in the case of HIV-1 protease, which would yield a greater enrichment than the ensemble. Unfortunately  
22 there is no clear way to a priori choose the best structure for docking.  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37

38 *Diversity of hits.* As noted above, not only are we interested in optimizing the recovery of actives in a  
39 virtual screening campaign, but equally important, we would also like to find a diverse set of hits on  
40 which to base further drug design efforts. The diversity of ligands is measured by number of clusters  
41 (~chemotypes) found at the 0.7 similarity level from a hierarchical clustering of the ligands, as described  
42 in methods. As with enrichment of HIV-1 protease hits, we see from Table 4 that the ensemble  
43 produces greater diversity than is achieved by docking into single structures, with the exception of the  
44 MD simulation. Again, one could achieve similar diversity from a single structure if one were fortunate  
45 enough to pick the best structure in which to dock.  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

*Summary of enrichment and diversity comparison for the three systems.* The results for all three systems studied here are summarized in Table 5, and the detailed tables for AR and CDK2 are given in the supplementary material (Tables S1 and S2). As we see from Table 5, in all cases the ensemble produces a better enrichment and greater diversity of hits than from the same number of compounds taken from docking into a single structure for the crystal structures. For example, the ensemble docking yields enrichment factors of 13.0, 6.3, and 9.1 for the crystal systems of AR, HIV, and CDK2, respectively while the average of the comparable enrichment factors for docking into single structures for the same systems is 9.7, 4.9, and 6.3 respectively. The diversity, as measured by the number of clusters/“chemotypes” recovered from the ensemble, is also superior (4, 6, and 11) as compared to the average crystal values (3.5, 5.0, and 7.0), respectively. Similar results are found from the structures derived from REMD, with the exception of AR. The MD results don’t follow this trend with both the AR and HIV-1 protease

systems showing greater or equal enrichment from docking into a single structure and the average diversity achieved is also greater in the top 4% of the hits from single structures, as opposed to the combined top 1% of each of the individual representative structures. In CDK2, docking into the ensemble again produces better enrichment and diversity than the single structures.

## Discussion and Conclusions.

We have shown that MD and REMD simulation techniques can be used to generate diverse structures that may be exploited to improve database enrichment and diversity in virtual screening campaigns as compared with the use of a single receptor structure. Furthermore, a procedure was described to select an ensemble of structures based on diversity of the binding site shapes. The ensemble selection is an unbiased procedure based on hierarchical clustering to select diverse binding site shapes.

1 The results presented here show that structures generated from MD and REMD simulations can be  
2 used in docking studies and provide significant enrichment and diversity of hits. In all three systems,  
3 docking into structures taken from MD and REMD yield good enrichment of actives, ranging from 3.8-  
4 fold improvement over random in the case of MD structures of CDK2 to 9.7-fold over random in AR,  
5 when analyzing the top 4% of the virtual screening hits. However, when many crystal structures are  
6 available, docking into an ensemble of binding site shape diverse crystal structures yields better  
7 enrichment and diversity of active ligands than docking into structures derived from simulations. Thus,  
8 one should use crystal structures if enough structures are available to achieve binding site diversity  
9 sufficient to accommodate the diversity of actives sought. The results from AR also confirm the  
10 conclusion drawn previously that docking into a small ensemble of representative crystal structures will  
11 yield better enrichment and diversity than docking into a single structure.  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27

28 The value of using REMD versus MD for the generation of receptor structures is not clear from the  
29 limited studies performed here. In HIV-1 protease, greater enrichment is achieved by the REMD  
30 ensemble docking whereas in CDK2 there is no difference in enrichment between MD and REMD. In  
31 AR, equal or better enrichment is achieved by docking into any of the single representative structures  
32 from the simulation techniques than exploiting the top 1% of each structure in the ensemble. The reason  
33 for this is the subject of further study.  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44

45 It was hypothesized that the relatively greater effectiveness of the crystal structures in the CDK2 and  
46 HIV-1 protease to yield enrichment in docking over the MD structures of AR was due to the limited  
47 sampling in the 1 ns portion of the MD trajectory exploited. Here, we employed REMD simulations to  
48 test this hypothesis. As seen from the resulting docking studies carried out in all three systems, the  
49 improved sampling afforded by the REMD did not translate into improved enrichment or diversity. The  
50 reason for this remains to be elucidated. The ligand binding site volume overlap matrices of the REMD  
51 with the crystal structures indicate that the simulations sample regions of LBS volumes which are not  
52  
53  
54  
55  
56  
57  
58  
59  
60

1 sampled by the crystal-ligand systems. This raises the question as to whether the REMD energy surface  
2 may not be adequately true to the actual energy surface of the protein ligand system, leading to slight  
3 artifactual excursions in the configuration space of the protein. That these deviations are not too extreme  
4 is demonstrated by the success of the structures in the trajectory to yield significant enrichment in  
5 docking, albeit not quite as competently as the crystal structures.  
6  
7  
8  
9  
10

11  
12  
13  
14 Perhaps a more plausible explanation is that binding site diversity alone is not a sufficient descriptor  
15 to determine the best structures to use for virtual screening. An extreme scenario can be constructed  
16 where a very diverse structure results from the binding site completely collapsing, yielding a “diverse”  
17 structure that would not accommodate any active compounds. While such an extreme situation was not  
18 encountered in the work presented here, more subtle situations may exist with side chains moving into  
19 positions that create diverse shapes that cannot accommodate active ligands. One possible solution  
20 would be to generate a diverse ensemble of structures under the constraint that certain other  
21 characteristics of the binding site are maintained, such as total volume, degree of enclosure, etc. Such a  
22 study is beyond the scope of the work presented here and will be the focus of future studies.  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37

38 This work was supported by the National Institutes of Health Grant R43 CA132538, and the National  
39 Science Foundation Grant CNS 0551500.  
40  
41  
42  
43  
44  
45

46 Acknowledgements. ATH would like to acknowledge Prof Harold Scheraga, a mentor and friend  
47 for many years.  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Supporting Information Description: This lists the PDB IDs for the AR crystal clustering, figures of the active site shapes for HIV and CDK2 from crystal, MD and REMD structures clustering, a dendrogram of the CDK2 crystal clustering, tables of detailed results for active ligand counts, enrichment factors and diversity estimates for docking to each individual crystal, MD and REMD cluster representative for AR, CDK2 and HIV and a dendrogram of HIV and CDK2 active ligand clustering using MACCs fingerprints. The supporting information includes additional figures and tables referenced in the paper. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141-151.
- (2) Schrodinger Software Suite. **2010**.
- (3) Osguthorpe, D. J.; Hagler, A. T.; Sherman, W. *Chem Biol Drug Des*, *in press*.
- (4) Osguthorpe, D. J.; Hagler, A. T. *Biochemistry* **2011**, *50*, 4105-4113.
- (5) Desmond Molecular Dynamics System. **2009**.
- (6) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.
- (7) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684-3690.
- (8) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117-1157.
- (9) Nose, S. *J. Chem. Phys.* **1984**, *81*, 511-519.
- (10) Hoover, W. G. *Physical Review A* **1985**, *31*, 1695-1697.
- (11) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604-1608.
- (12) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345-354.
- (13) Garcia, A. E.; Onuchic, J. N. *Proc Natl Acad Sci U S A* **2003**, *100*, 13898-13903.
- (14) Metropolis, N.; Ulam, S. *J Am Stat Assoc* **1949**, *44*, 335-341.
- (15) Patriksson, A.; van der Spoel, D. *PCCP* **2008**, *10*, 2073-2077.

- 1 (16) Tom, H. *Chem Biol Drug Des* **2007**, *69*, 146-148.
- 2
- 3 (17) Halgren, T. A. *J. Chem. Inf. Model.* **2009**, *49*, 377-389.
- 4
- 5
- 6 (18) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J Med Chem* **2006**, *49*, 6789-6801.
- 7
- 8
- 9
- 10 (19) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.;
- 11 Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J Med*
- 12 *Chem* **2004**, *47*, 1739-1749.
- 13
- 14
- 15
- 16
- 17 (20) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.;
- 18 Banks, J. L. *J Med Chem* **2004**, *47*, 1750-1759.
- 19
- 20
- 21
- 22
- 23 (21) MDL Information Systems "MACCS keys."
- 24
- 25
- 26 (22) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. *J. Chem. Inf. Model.* **2010**, *50*, 771-
- 27
- 28 784.
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

Tables

Table 1. Normalized overlap volume matrix of the four crystal cluster representatives for AR, the four MD cluster representatives for AR and the four REMD cluster representatives for AR 310 K. Cross normalized overlap volumes between the crystal, MD and REMD shapes shaded green

Crystal														
	1T74	2AMA	2AXA	2HVC										
1T74	1.00	0.69	0.45	0.66										
2AMA	0.69	1.00	0.56	0.57										
2AXA	0.45	0.56	1.00	0.51										
2HVC	0.66	0.57	0.51	1.00										
MD					e	f	g	h						
e	0.70	0.57	0.41	0.57	1.00	0.71	0.59	0.60	0.63	0.57	0.38	0.50		
f	0.71	0.59	0.46	0.60	0.71	1.00	0.55	0.61	0.64	0.59	0.38	0.53		
g	0.57	0.57	0.48	0.46	0.59	0.55	1.00	0.69	0.60	0.53	0.45	0.44		
h	0.56	0.55	0.49	0.46	0.66	0.61	0.69	1.00	0.64	0.55	0.42	0.41		
REMD									i	j	k	l		
i	0.63	0.57	0.38	0.50	0.64	0.64	0.57	0.54	1.00	0.70	0.60	0.64		
j	0.70	0.59	0.38	0.53	0.69	0.70	0.54	0.53	0.70	1.00	0.58	0.55		
k	0.56	0.53	0.45	0.44	0.57	0.57	0.59	0.56	0.60	0.58	1.00	0.50		
l	0.51	0.42	0.29	0.41	0.53	0.51	0.50	0.41	0.64	0.55	0.50	1.00		

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Table 2. Normalized overlap volume matrix of the four crystal cluster representatives for HIV, the four MD cluster representatives for HIV-1 protease and the four REMD cluster representatives for HIV-1 protease 310 K. Cross normalized overlap volumes between the crystal, MD and REMD shapes shaded green.

Crystal														
	1EBZ	1HVS	1XL2	3AID										
1EBZ	1.00	0.56	0.35	0.39										
1HVS	0.56	1.00	0.38	0.40										
1XL2	0.35	0.38	1.00	0.37										
3AID	0.39	0.40	0.37	1.00										
MD					e	f	g	h						
e	0.29	0.29	0.25	0.30	1.00	0.38	0.43	0.41						
f	0.38	0.33	0.32	0.30	0.38	1.00	0.42	0.46						
g	0.26	0.27	0.23	0.28	0.43	0.42	1.00	0.50						
h	0.29	0.29	0.25	0.27	0.41	0.46	0.50	1.00						
REMD									i	j	k	l		
i	0.36	0.37	0.30	0.32	0.35	0.36	0.35	0.36	1.00	0.24	0.33	0.26		
j	0.24	0.22	0.24	0.26	0.34	0.37	0.38	0.34	0.24	1.00	0.37	0.35		
k	0.37	0.38	0.28	0.39	0.34	0.28	0.33	0.29	0.33	0.37	1.00	0.37		
l	0.15	0.19	0.24	0.24	0.27	0.34	0.29	0.30	0.26	0.35	0.37	1.00		

Table 3. Normalized overlap volume matrix of the four crystal cluster representatives for CDK2, the four MD cluster representatives for CDK2 and the four REMD cluster representatives for CDK2 310 K. Cross normalized overlap volumes between the crystal, MD and REMD shapes shaded green.

Crystal														
	1OI9	2BHE	1GZ8	1W0X										
1OI9	1.00	0.46	0.40	0.49										
2BHE	0.46	1.00	0.38	0.52										
1GZ8	0.40	0.38	1.00	0.50										
1W0X	0.49	0.52	0.50	1.00										
MD					e	f	g	h						
e	0.36	0.32	0.45	0.44	1.00	0.37	0.31	0.29						
f	0.32	0.26	0.38	0.35	0.37	1.00	0.41	0.39						
g	0.28	0.27	0.30	0.31	0.31	0.41	1.00	0.43						
h	0.52	0.41	0.37	0.45	0.29	0.39	0.43	1.00						
REMD									i	j	k	l		
i	0.51	0.43	0.33	0.42	0.30	0.35	0.37	0.50	1.00	0.49	0.41	0.38		
j	0.47	0.39	0.40	0.42	0.31	0.37	0.28	0.51	0.49	1.00	0.37	0.40		
k	0.44	0.43	0.44	0.55	0.49	0.35	0.31	0.43	0.41	0.37	1.00	0.35		
l	0.37	0.28	0.46	0.37	0.46	0.47	0.40	0.38	0.38	0.40	0.35	1.00		

Table 4. HIV-1 protease Active Ligands and Decoy Ligands Docking to 4 Representative Structures from Crystal, MD and REMD. Active ligand counts, enrichment factors and Diversity for each structure at the 1% and 4% level							
		Count of Active Ligands		Enrichment Factor		Diversity of Actives	
Cluster Representative		Top 19 (1%) <sup>†</sup>	Top 76 (4%)	Top 19 (1%)	Top 76 (4%)	Top 19 (1%)	Top 76 (4%)
Crystal	1	6 (6)	10	12	4.8	3	5
	2	10 (5)	16	19	7.7	5	6
	3	3 (0)	6	6	2.9	2	4
	4	7 (2)	9	13	4.3	3	5
Ensemble		13	10.3 <sup>††</sup>	6.3	4.9	6	5.0
MD	1	8 (2)	13	15	6.3	4	6
	2	2 (0)	8	3.8	3.8	1	3
	3	6 (1)	12	12	5.8	1	4
	4	7 (7)	8	13	3.8	3	3
Ensemble		10	10.3	4.8	4.9	2.3	4.0
REMD	1	6 (6)	11	12	5.3	1	3
	2	8 (6)	17	15	8.2	5	6
	3	2 (0)	6	3.9	2.9	2	2
	4	4 (1)	8	7.7	3.8	2	3
Ensemble		13	10.5	6.3	5.0	5	3.5
<sup>†</sup> Additional unique ligand count shown in brackets – these are the ligands of the current structure that are different from any active ligand of structures before this structure							
<sup>††</sup> This is average over the active ligands recovered in the top 4% of hits							



Table 5. Comparison of ensemble 1% results with average of 4% results for Crystal, MD and REMD studies of AR, HIV-1 protease and CDK2

		Count of Active Ligands		Enrichment Factor		Diversity of Actives	
System		Ensemble 1% <sup>a</sup>	Average 4% <sup>b</sup>	Ensemble 1% <sup>c</sup>	Average 4% <sup>d</sup>	Ensemble 1% <sup>e</sup>	Average 4% <sup>f</sup>
AR	Crystal	39	29	13	9.7	4	3.5
	MD	25	27	8.5	9.0	3	2.3
	REMD	20	29	6.8	9.7	2	2.3
HIV	Crystal	13	10.3	6.3	4.9	6	5.0
	MD	10	10.3	4.8	4.9	2	4.0
	REMD	13	10.5	6.3	5.0	5	3.5
CDK2	Crystal	18	12	9.1	6.3	11	7.0
	MD	12	7.8	6.1	3.9	8	5.3
	REMD	12	7.5	6.1	3.8	8	5.5

<sup>a</sup>. Count of unique active ligands from all 4 Structures at the 1% level

<sup>b</sup>. Average of active ligand count for the 4 Structures at the 4% level

<sup>c</sup>. EF computed using the count of unique active ligands from all 4 structures at the 1% level

<sup>d</sup>. Average of the EF computed at the 4% level for each structure

<sup>e</sup>. Diversity computed using the unique active ligands from all 4 structures at the 1% level

<sup>f</sup>. Average of the Diversity computed at the 4% level for each structure

## Figure Captions

Figure 1. Dendrogram of clustering the 74 active ligands of the AR DUD dataset by their MACCS fingerprints, defining clusters at the 0.7 level and using the same color for the members of each cluster and blue for singleton clusters. The x-axis is the ZINC ID numbers.

Figure 2. Trajectory of residue-residue distances spanning the binding sites for the AR, HIV-1 protease and CDK2 MD (red) and REMD (green) simulations. The AR trajectory is characterized by the distance between the Gln 711 side chain amide nitrogen and the Met 895 sulfur. The HIV-1 protease trajectory corresponds to the Val 82 – Val 182 side chain distance, the CDK2 trajectory by the Tyr 15 - Asn 132 side chain distance.

Figure 3. Active site shapes for the androgen receptor from crystal structures, MD and REMD. The four crystal structure cluster representatives of AR are given in panels a-d: The four cluster representatives from MD are given in structures e-h, while the four cluster representatives from REMD are given in i-l. Helix 3 of AR, which runs roughly vertically in front of the LBS, has been removed from view for clarity.

Figures

Figure 1

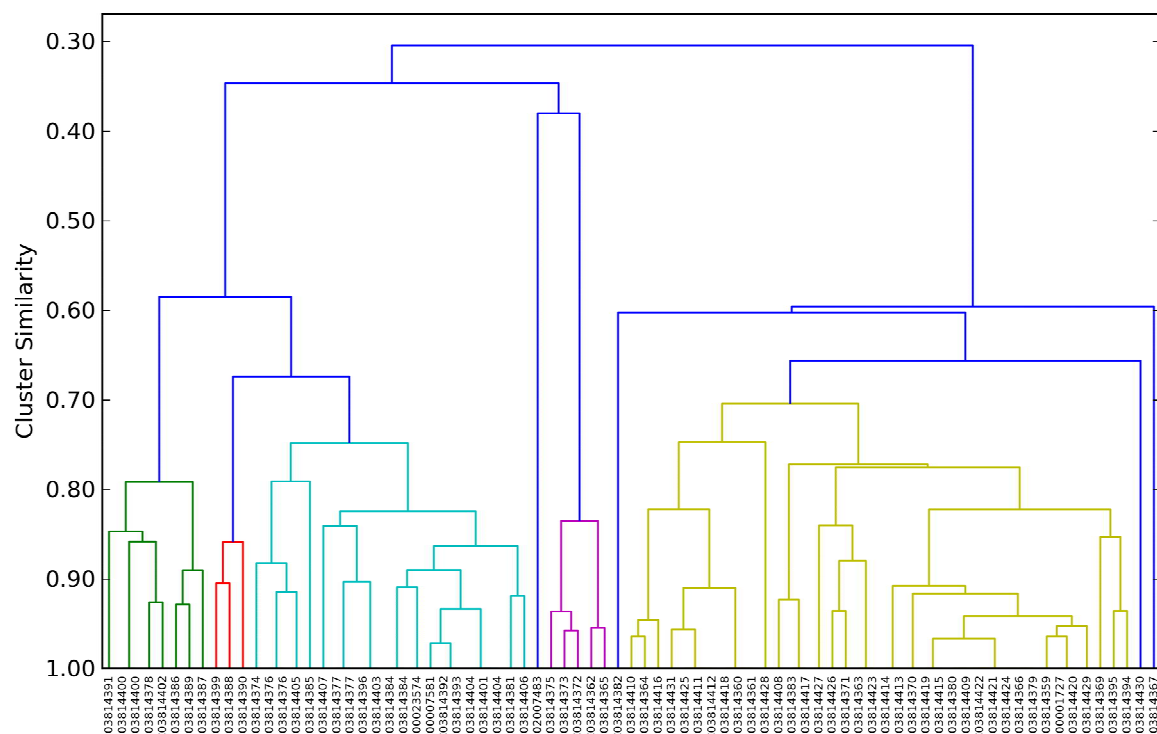


Figure 2

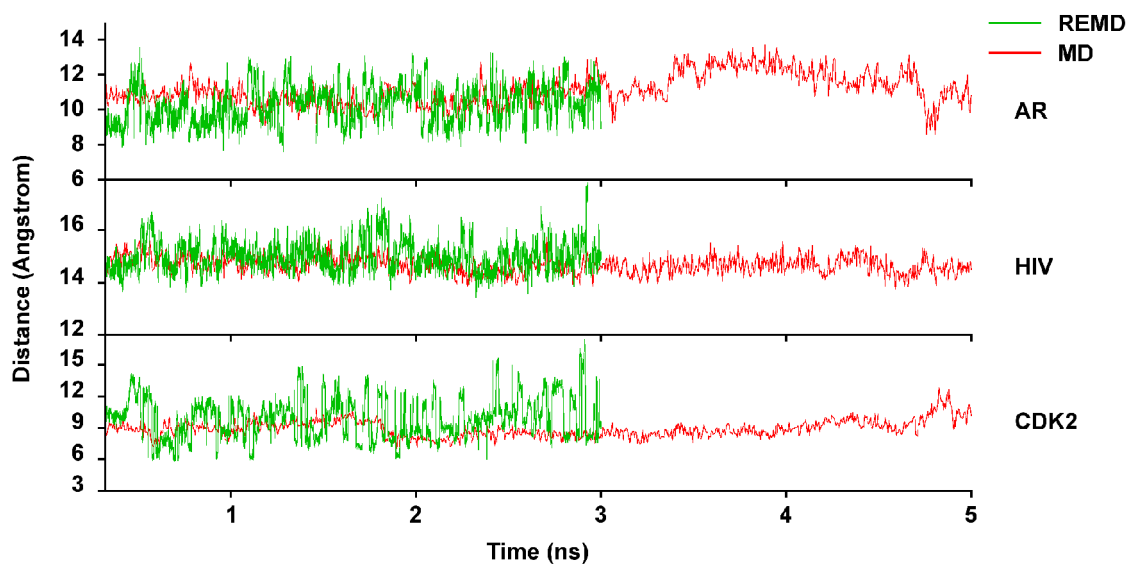
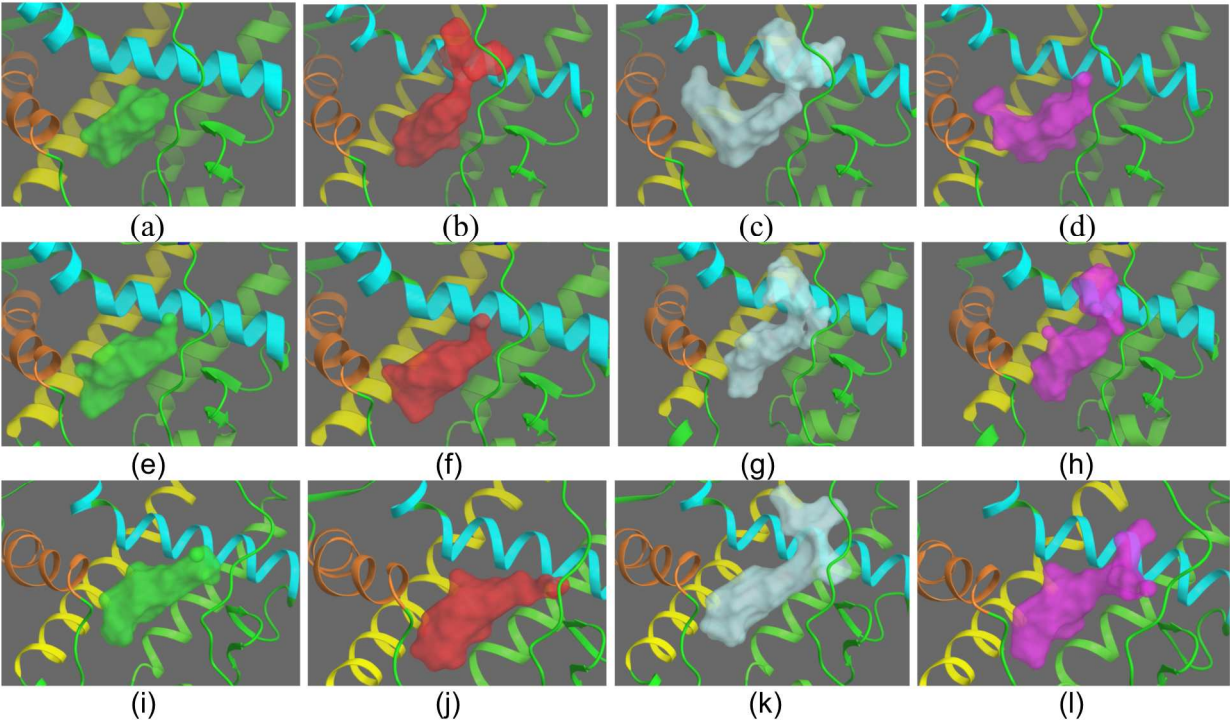


Figure 3



# The Journal of Physiology

1

2

3

4

5

6

Paragon Plus Environment