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

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

1. Shifa Biomedical, 1 Great Valley Parkway, Suite 8, Malvern, Pennsylvania 19355, United States
2. Schrödinger, Inc., 120 West 45th Street, 17th 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

PDB IDs for AR Crystal Clustering.

1E3G, 1T5Z, 1T63, 1T65, 1T73, 1T74, 1T76, 1T79, 1T7F, 1T7M, 1T7R, 1T7T, 1XJ7, 1XOW, 1XQ3, 2AM9, 2AMA, 2AMB, 2AO6, 2AX9, 2AXA, 2HVC, 2PIO, 2PIP, 2PIQ, 2PIR, 2PIT, 2PIU, 2PIV, 2PIW, 2PIX, 2PKL, 2PNU, 2Q7I, 2Q7J, 2QPY, 2Z4J, 3B5R, 3B65, 3B66, 3B67, 3B68, 3L3X.

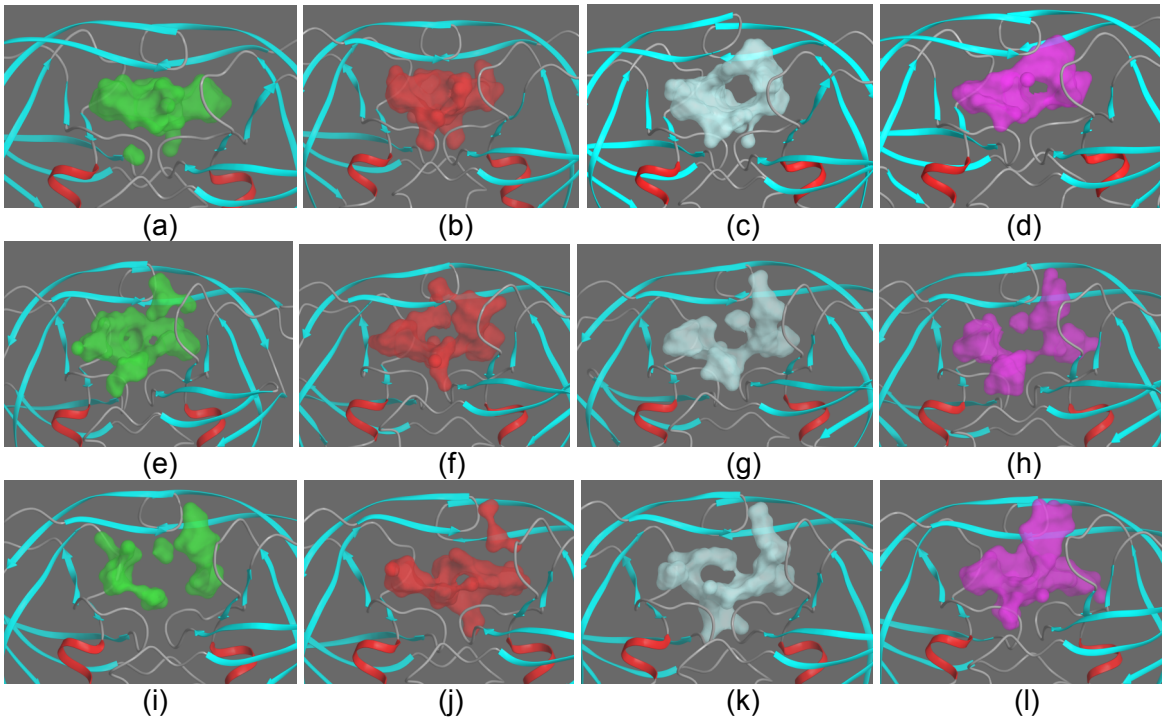

Figure S1. Active site shapes for the cluster representatives of HIV protease: (a,b,c,d) , Active site shapes for the four crystal structure cluster representatives of HIV protease, PDB IDs 1EBZ, 1HVS, 1XL2 and 3AID respectively; (e,f,g,h) ,The four cluster representatives from analysis of 800 MD structures of 1EBZ HIV structure. There are 406, 54, 114 and 227 structures in the clusters represented by the representative structures e-h respectively; (i,j,k,l), The four cluster representatives from clustering 1000 structures derived from 2ns of REMD. There are 496, 321, 117 and 67 structures in the clusters represented by (i), (j), (k) and (l) respectively.

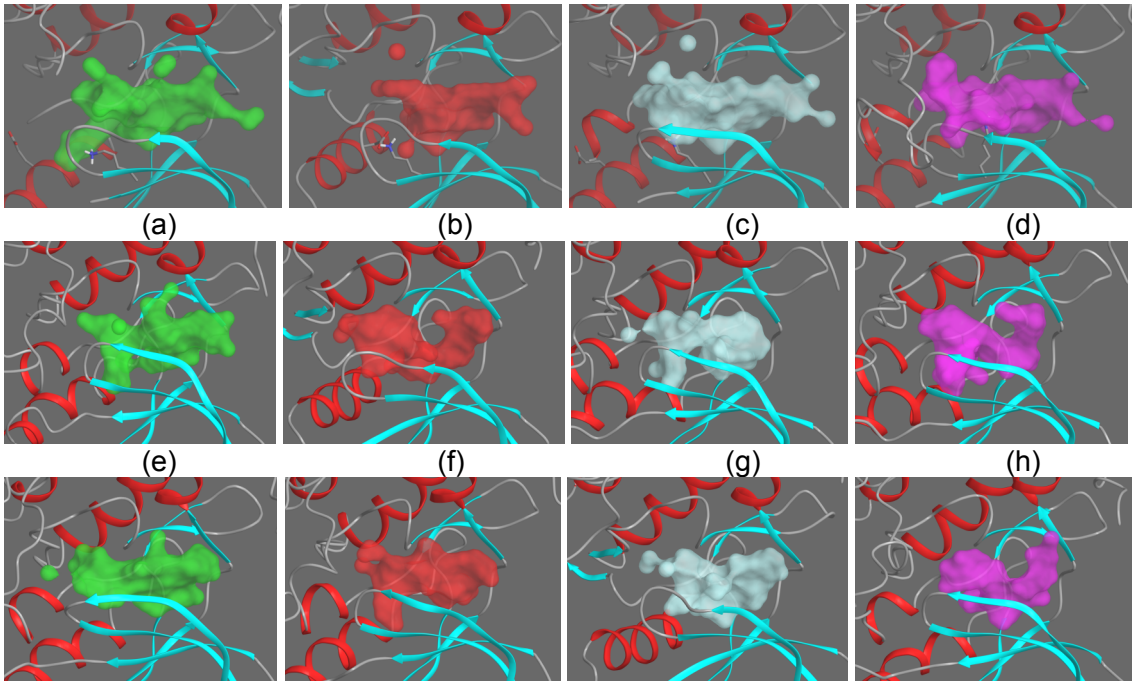

Figure S2. Active site shapes for cluster representatives of CDK2: (a,b,c,d), Active site shapes for the four crystal structure cluster representatives of CDK2, PDB IDs 2BHE, 1OI9, 1W0X and1GZ8. There are 13, 27, 38 and 14 structures in the clusters represented by (a), (b), (c) and (d) respectively; (e,f,g,h), The four cluster representatives from analysis of a combined 800 MD structures of non-cyclin bound (1W0X) and 800 MD structures of cyclin bound (1OI9) CDK2 crystal structures. There are 676, 803, 67 and 56 structures in the clusters represented by (e), (f), (g) and (h) respectively; (I,j,k,l): The four cluster representatives from analysis of a combined 1000 REMD structures of non-cyclin bound (1W0X) and 1000 REMD structures of cyclin bound (1OI9) CDK2 crystal structures. There are 331, 222, 1104 and 345 structures in the clusters represented by (j), (j), (k) and (l) respectively.

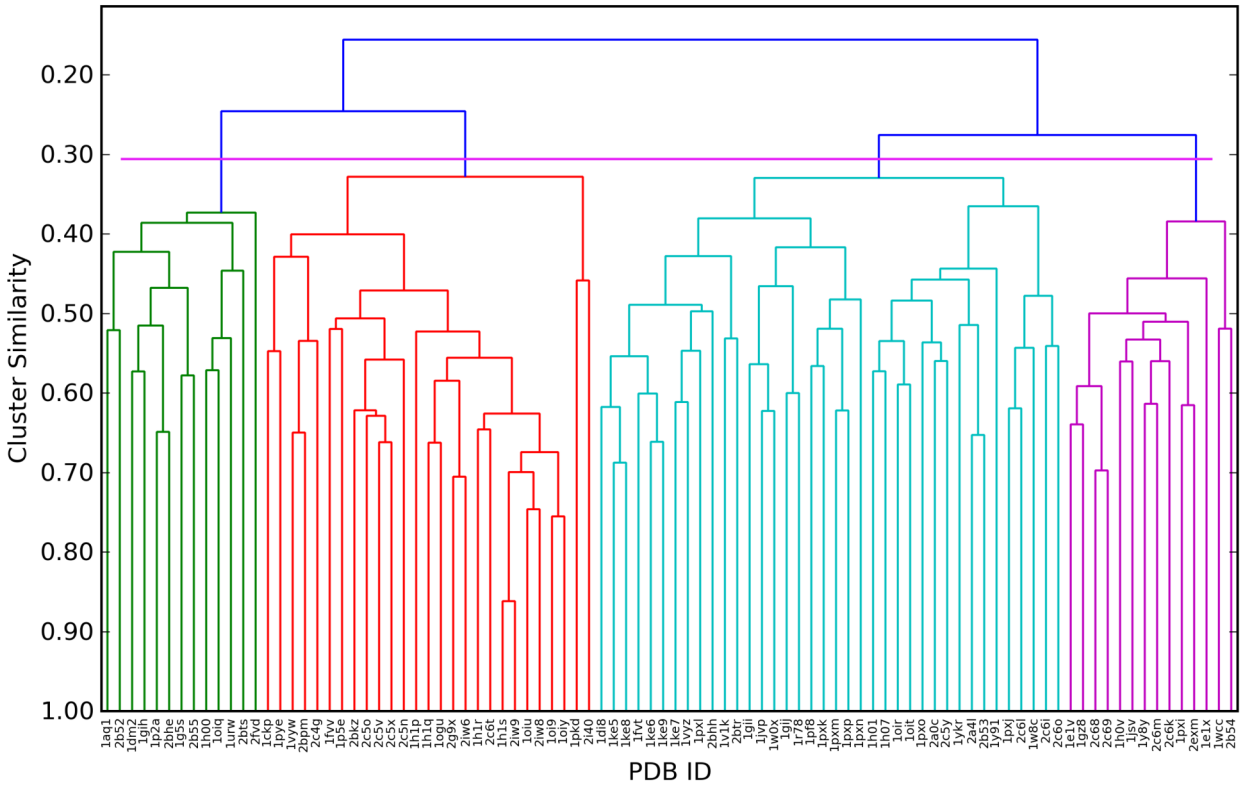

Figure S3. Dendrogram of Ligand Binding Site Volume (LBSV) Clustering of 92 structures of CDK2 coloring at the 4-cluster level. For each of the 4 clusters the cluster representatives were selected as described in methods, and used in the docking studies. The magenta line shows the similarity level required to divide structures into four clusters. The red cluster contains all cyclin bound CDK2 structures plus an additional two non-cyclin bound structures. The other three clusters contain all non-cyclin bound structures.

Table S1. Docking of AR ligands to 4 Representative Structures from Crystal, MD and REMD. Active ligand counts, enrichment factors and diversity estimate are given for each structure at the 1% and 4% level*

		Count of Active Ligands		Enrichment Factor		Diversity of Actives	
Cluster Representative		Top 27 (1%)	Top 108 (4%)	Top 27 (1%)	Top 108 (4%)	Top 27 (1%)	Top 108 (4%)
Crystal	1	12 (12)	25	16	8.5	1	3
	2	14 (5)	23	19	7.8	1	2
	3	15 (11)	28	20	9.5	5	5
	4	22 (11)	39	30	13	4	4
Ensemble		39	29	13	9.7	2	2.3
MD	1	18 (18)	25	24	8.5	1	2
	2	20 (4)	26	27	8.8	1	1
	3	20 (3)	32	26	10	3	4
	4	15 (0)	25	20	8.5	2	2
Ensemble		25	27	8.5	9.0	3	2.3
REMD	1	12 (12)	36	16	12	3	4
	2	14 (7)	22	19	7.4	1	1
	3	11 (0)	19	15	6.4	1	1
	4	15 (1)	37	20	13	1	3
Ensemble		20	29	6.8	9.7	2	2.3

* 74 Ligands, 2628 Decoys, Totals 2702 Total Diversity 9

† 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

Table S2. . Docking of CDK2 ligands to 4 Representative Structures from Crystal, MD and REMD. Active ligand counts, enrichment factors and diversity estimate are given for each structure at the 1% and 4% level*

		Count of Active Ligands		Enrichment Factor		Diversity of Actives	
Cluster Representative		Top 18 (1%)	Top 72 (4%)	Top 18 (1%)	Top 72 (4%)	Top 18 (1%)	Top 72 (4%)
Crystal	1	7 (7)	11	14	5.6	5	6
	2	10 (6)	18	20	9.1	6	11
	3	3 (1)	10	6.1	5.1	2	6
	4	4 (4)	11	8.1	5.6	3	5
Ensemble		18	12	9.1	6.3	11	7.0
MD	1	4 (4)	9	8.1	4.6	3	7
	2	5 (5)	14	10.2	7.1	5	9
	3	2 (1)	4	4.1	2.0	2	2
	4	4 (2)	4	8.1	2.0	3	3
Ensemble		12	7.8	6.1	3.9	8	5.3
REMD	1	5 (5)	10	10.2	5.1	3	7
	2	1 (1)	4	2.0	2.0	1	4
	3	2 (2)	7	4.1	3.6	2	5
	4	5 (4)	9	10.2	4.6	3	6
Ensemble		12	7.5	6.1	3.8	8	5.5

* 50 Ligands, 1779 Decoys, Totals 1829 Total Diversity 26

† 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

Table S3. Docking of HIV Protease ligands to 4 Representative Structures from Crystal, MD and REMD. Active ligand counts, enrichment factors and diversity estimate are given for each structure at the 1% and 4% level*

		Count of Active Ligands		Enrichment Factor		Diversity of Actives	
Cluster Representative		Top 19 (1%)†	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
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
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

* 53 Ligands, 1885 Decoys, Totals 1938 Total Diversity 13

† 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

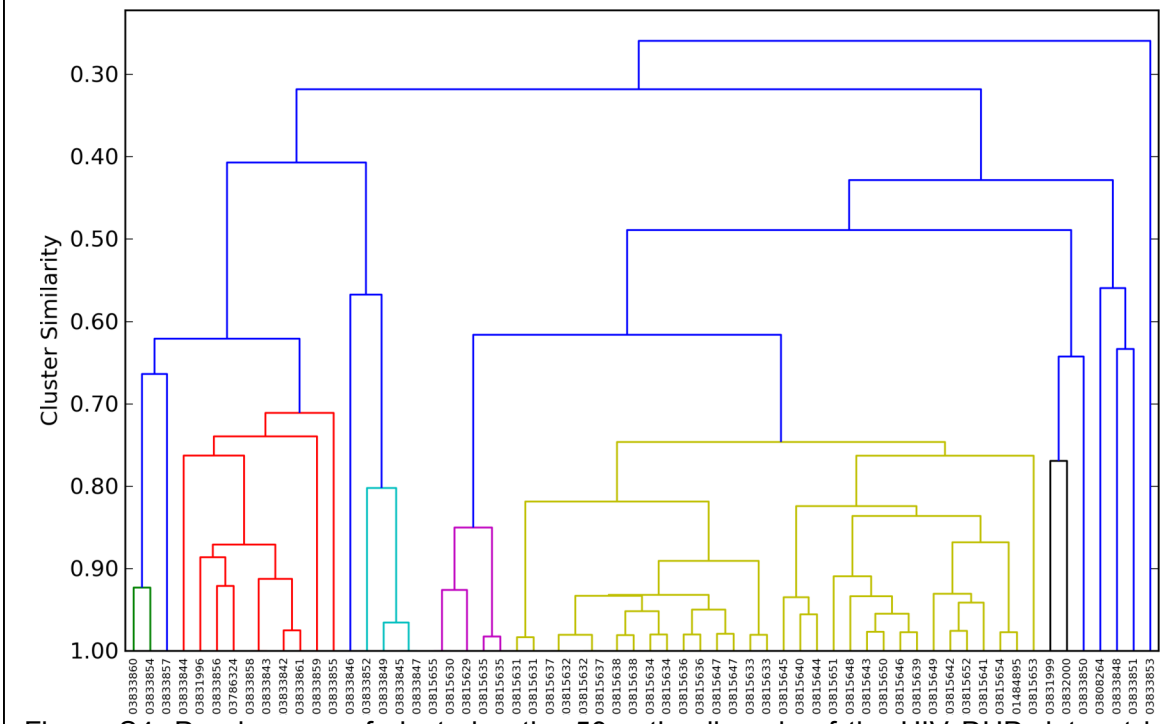

Figure S4. Dendrogram of clustering the 53 active ligands of the HIV DUD dataset by their MACCs fingerprints. Clusters are defined at the 0.7 similarity level and color coded. Blue is used for the singleton clusters. The x-axis is the ZINC ID numbers. The total number of clusters at the 0.7 level is 13.

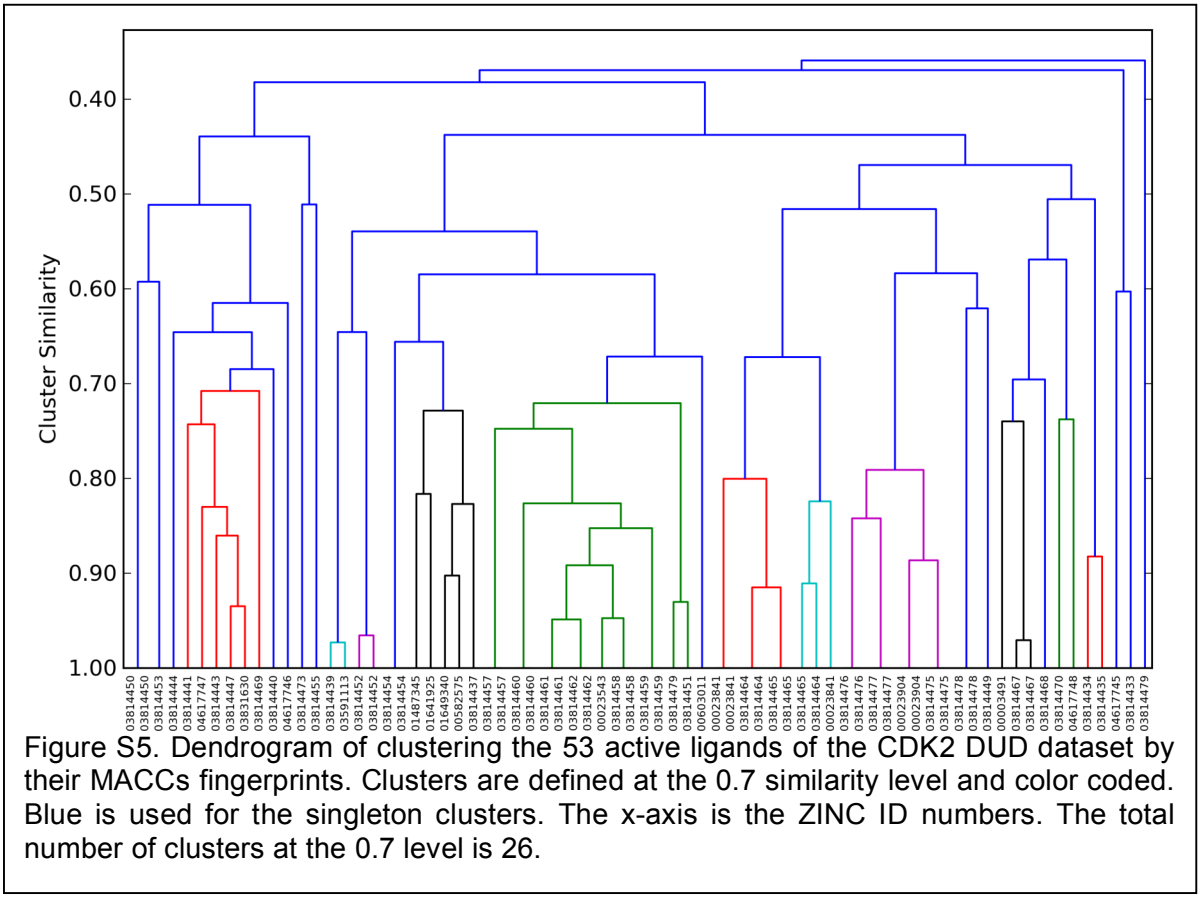

Figure S5. Dendrogram of clustering the 53 active ligands of the CDK2 DUD dataset by their MACCs fingerprints. Clusters are defined at the 0.7 similarity level and color coded. Blue is used for the singleton clusters. The x-axis is the ZINC ID numbers. The total number of clusters at the 0.7 level is 26.