

Pre-docking filter for protein and ligand 3D structures

Alisa Wilantho¹, Sissades Tongsim¹ and Ekachai Jenwitheesuk^{1,*}

¹National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, 113 Thailand Science Park, Phahonyothin Road, Klong 1, Klongluang, Pathumtani 12120, Thailand;
Ekachai Jenwitheesuk* - Email : ekachai@biotec.or.th; Phone: 66 2564 6700; Fax: 66 2564 6701; * Corresponding author

received November 15, 2008; accepted December 01, 2008; published December 31, 2008

Abstract:

Virtual drug screening using protein-ligand docking techniques is a time-consuming process, which requires high computational power for binding affinity calculation. There are millions of chemical compounds available for docking. Eliminating compounds that are unlikely to exhibit high binding affinity from the screening set should speed-up the virtual drug screening procedure. We performed docking of 6353 ligands against twenty-one protein X-ray crystal structures. The docked ligands were ranked according to their calculated binding affinities, from which the top five hundred and the bottom five hundred were selected. We found that the volume and number of rotatable bonds of the top five hundred docked ligands are similar to those found in the crystal structures and corresponded with the volume of the binding sites. In contrast, the bottom five hundred set contains ligands that are either too large to enter the binding site, or too small to bind with high specificity and affinity to the binding site. A pre-docking filter that takes into account shapes and volumes of the binding sites as well as ligand volumes and flexibilities can filter out low binding affinity ligands from the screening sets. Thus, the virtual drug screening procedure speed is increased.

Keywords: virtual drug screening; ligand volume; protein binding site; docking filter; binding affinity

Background:

Virtual screening techniques are becoming increasingly more important in drug discovery. A popular method for virtual screening is molecular docking [1, 2], which selects small-molecule structures from databases such as ChemBank [3], ChemPDB [4], KEGG [5], and NCI [6] and docks them into the protein binding site [7]. These processes involve the prediction of binding energies and analysis of molecular binding modes, which are time consuming and computationally expensive. The two-dimensional (2D) fingerprint technique, a virtual screening method which measures the structural similarity of molecules has been developed to address the above problems [8]. The similarity search is based upon the "similar property principle", which states that molecules that are structurally similar are likely to have similar properties [9]. This technique uses a ligand with known chemical properties, inhibitory activities, or binding modes for a target of interest as a reference for searching similar ligands in the database regardless of the shape and size of the protein binding site. The accuracy of this method depends on which similarity coefficient is used [10-12], and the Tanimoto coefficient is most popularly employed [13]. Based on the "lock-and-key" principle, we propose a novel pre-docking procedure that matches the sizes of the ligand with the protein binding site, and optimizes the grid-box size before docking. This simple procedure dramatically reduces the size of screening ligand sets, significantly reducing time and effort required for virtual drug screening.

Methodology:

Preparation of ligand and protein structures for docking

Preparation of ligand structures

The ligands used in this study were downloaded from Ligand.Info Meta-Database [14]. Ligands set-1 consisted of 2344 structures from ChemBank and set-2 consisted of 4009 structures from ChemPDB. The downloaded ligands in the SDF format were first converted to the Protein Data Bank (PDB) format using Open Babel [15]. The Gasteiger charges and rotatable bonds were then assigned to the PDB ligands using AutoDockTool [16]. All rotatable bonds were allowed to move freely.

Preparation of protein structures

Twenty-one protein X-ray crystal structures from the Protein Data Bank [17] were downloaded. The proteins and their PDB structure identifiers (PDB ID) are given in Table 1 (supplementary material). Of the twenty-one protein structures, sixteen have co-crystallized ligands (X-ray ligand) in the binding site. The ligand contained in each protein structure was removed from the binding site and saved to a new file. The missing atoms in each protein structure were searched for and fixed using SwissPDB [18]. The Gasteiger charges and the solvation term were then added to the protein structure using the AutoDockTool.

Ligand SMILES string similarity search

The ligands extracted from the X-ray crystal structure obtained from the previous step were converted to the SMILES string format, and used as an input for similarity search against the ligands in the ChemBank and the ChemPDB sets in step of the Preparation of ligand structures using Tanimoto coefficient cutoffs of 0.5, 0.6, and 0.7, respectively.

Calculation of ligand molecular volume

The volumes of the ligands in the screening set were calculated using Mol_Volume version 1.0 [19]. The van der Waals radii value for each atom type was derived from the CHARMM 22 force field. The radius of the spherical probe (R_PROBE) was set to 2.0 Å, and the GRID_STEP was set to 0.5 Å. The volumes of the ligands extracted from the X-ray crystal structures were calculated using the same protocol. The extracted ligand name (X-ray ligand) and its calculated volumes are shown in Table 2 (supplementary material).

Calculation of protein binding site molecular volume

Protein binding site volumes were calculated using the CASTp server (<http://sts-fw.bioengr.uic.edu/castp>) [20]. The solvent probe radius used for volume calculation was 1.4 Å. CASTp identifies all surface pockets with the chosen volume values, and then displays them on the computer screen. Pockets calculated by CASTp that matched the pocket resolved by crystallography were selected, and the volume of that calculated pocket was taken as the volume of the protein binding site.

Protein-ligand docking

Grid-box generation

The grid parameter file of each protein was generated using AutoDockTool. A grid-box was generated that was large enough to cover the entire protein binding site and accommodate all ligands to move freely. The number of grid points in x, y, and z-axes were 60×60×60. The distance between two connecting grid points was 0.375 Å. The center of the ligand in the X-ray crystal structure was used as the center of the grid-box. For protein structures that do not have ligands in the binding site, the center of the binding site was estimated from the structure and taken as the center of the grid-box.

Ligand docking

AutoDock4 and a Lamarckian Genetic Algorithm (LGA) [21] were used for protein-fixed ligand-flexible docking calculations. Ten search attempts (ga_run parameter) were performed for each ligand. The maximum number of energy evaluations before the termination of LGA run was 2500000 and the maximum number of generations of the LGA run before termination was 27000. Other docking parameters were set to the software's default values. After docking, the ligands were ranked according to their protein-ligand affinity (calculated inhibitory constant, K_i).

Discussion:

Size and shape of protein binding site

In this study, we categorized protein binding sites according to their sizes and shapes. Protein binding sites were classified as small (less than 1200 Å³) or large

(greater than or equals to 1200 Å³). Protein binding site shapes were classified as either simple or complex. The protein binding site classifications are shown in Table 1 (supplementary material). Sixteen protein structures had co-crystallized ligands bound in the binding site. The calculated volumes show that the majority of ligands are larger (305–5922 Å³) than the binding sites (1040–2690 Å³) in particular those in the small binding site group. However, the average volumes of the ligands (1684 Å³) and the binding sites (1638 Å³) are very similar. The protein binding site typically accommodates 50–70% of the ligand, with the remainder of the ligand occupying pockets adjacent to the binding site. For example, 50% of the GDP ligand (volume = 1460 Å³) was contained within the small binding pocket (volume = 594 Å³) of the “Filamenting temperature-sensitive mutant Z” protein (PDB ID: 1RQ7), while the rest of GDP ligand occupied pockets close to or floating over the binding site. This suggests that the optimal ligand size can potentially exceed the binding site volume.

X-ray ligand docking and ranking

To verify the docking procedure utilized in this work, we re-docked the original X-ray ligand back to its corresponding protein binding site. The X-ray ligands along with all other ligands in the screening set were ranked according to the calculated K_i . The X-ray ligands were ranked in the top ten percentiles and were also able to move back to the original positions with the root mean square deviations of less than 3 Å.

Docked ligand size and flexibility

The top 500 and the bottom 500 ligands ranked according to the K_i value for each protein were selected for further analysis. A scatter plot of the molecular volumes and the number of active torsion bonds for these ligands is shown in **Figure 1**. The top 500 ligands are clearly coincident with the sixteen X-ray ligands, which occupy volumes of 800–2800 Å³, whereas the bottom 500 ligands occupy volumes outside this range, with 95% much smaller (300–900 Å³). There does not appear to be any correlation between the number of active torsion bonds and calculated K_i ; however, the majority of the top 500 ligands have twenty or fewer active torsion bonds. These data suggest that ligands with high binding affinity are constrained by their size (volume 800–2800 Å³) and flexibility (20 or fewer active torsion bonds). For untested ligands, these parameters could be useful to prioritize docking calculations, so that priority is given to ligands of optimal size and flexibility.

Optimal size of the grid-box

In this study, a very large grid-box (10830 Å³, 22.125 Å on each side) was used because we wanted to ensure that the grid-box could cover the entire binding site, and that all ligands in the screening sets had enough space to enter and move freely in the grid-box. The volume of the grid-box was 10830 Å³ while the volume of the largest protein binding site was only 5921.8 Å³ (PDB ID: 1N8W). We hypothesized that using a very large grid-box would allow the binding of some ligands to extend beyond the actual binding site, with non-specific binding into adjacent pockets. We tested this hypothesis by generating minimal

grid-boxes that perfectly encompassed the entire binding site for each protein. The box sizes and their dimensions

are shown in Table 3 (supplementary material). The top 500 ligands were left on protein at the docked positions.

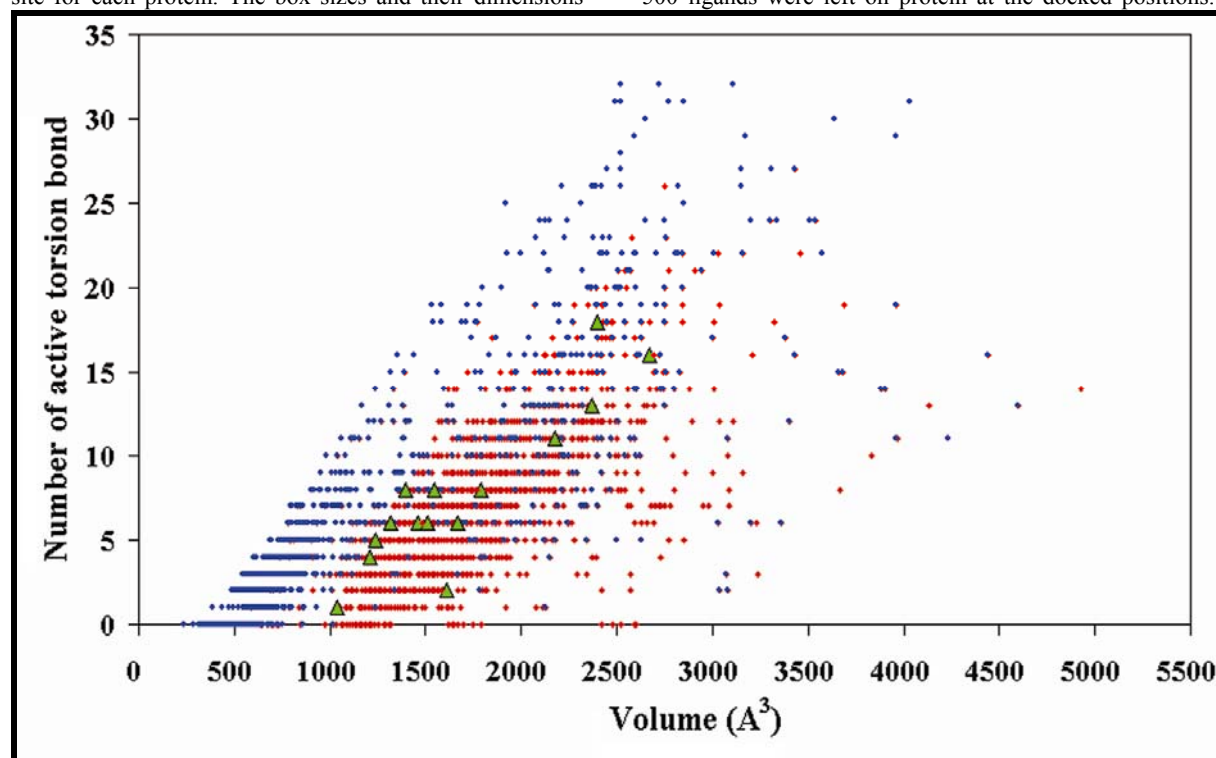


Figure 1: The volume and the number of active torsion bonds of the top 500 (red) and the bottom 500 docked ligands (blue) ranked according to the calculated inhibitory constant (K_i). Of the twenty-one protein X-ray crystal structures used in this study, sixteen structures had ligand bound in the binding site. The top 500 ligands generally had structural profiles in terms of volume and number of active torsion bond similar to those of the X-ray ligands (green) while the bottom 500 ligands were, on average, 800-900 Å³ smaller than the X-ray ligands.

Filtering of the top 500 docked ligands was performed to test how ligands occupy space beyond the protein binding site. Six thresholds of decreasing stringency, allowing progressively more of the ligand atoms to be outside the minimal grid-box were used (Table 4 in supplementary material). It is clear that on average, the top 500 ligands cannot fit entirely within the minimal grid-boxes extending outside them, since 10.8% of the ligands were rejected, even when a very relaxed 30% threshold was employed (Table 4 in supplementary material). Visual inspection of the docked structures revealed that the protein binding sites contain at least one opening space, and parts of the docked ligands were always outside of the minimal box on this side. The rejected ligands might be either too large or too long to fit entirely within the minimal box, or their chemical properties may not match perfectly well with the binding pocket so that parts of their structures bind preferentially with adjacent pockets.

Reducing the grid-box size would significantly reduce CPU time for docking calculation, an important consideration for drug-discovery when potentially millions of compounds are screened. However, it is clear from our data that this would also increase the false negative rate, leading to some high binding-affinity ligands to be missed. These false-negatives would likely include molecules with long linear shapes, or with branches which extend beyond the target binding site and bind to adjacent pockets, in

particular on the opening space side (see above). We propose that the optimal grid-box size allows approximately two-thirds of ligand molecule to occupy the target binding site, with the remaining one third able to bind with adjacent pockets. Grid-boxes of this size provide the optimal balance between the number of screening ligands and the CPU time required for docking.

SMILES strings similarity of the docked ligands and the X-ray ligands

We further explored whether the top 500 and the bottom 500 ligands docked on each protein were chemically similar to the X-ray ligand extracted from the protein-ligand co-crystal structures using the SMILES strings similarity search. The results show that, in general, more of the top 500 ligands matched with the X-ray ligands than of the bottom 500 ligands (Table 5 in supplementary material). On the other hand, even at a Tanimoto coefficient of 0.5, only thirteen ligands in the top500 list matched with the X-ray ligand. This indicates that most of the potential hits were chemically dissimilar to the X-ray ligands, yet similar in size as discussed above. Conversely, at a Tanimoto coefficient of 0.5, six out of the bottom 500 ligands matched with the X-ray ligand, suggesting that these ligands although similar to the X-ray ligand have chemical properties that are unfavorable to interactions with the binding site.

Conclusion:

There are millions of ligand structures currently available in public databases. Virtual screening of these ligands against a protein target using protein-ligand docking methods requires lengthy calculations on a high performance computer. This is a major obstacle that prevents several research groups in academia, especially those with limited computer resources to conduct research in this field. To overcome this problem, several computational techniques have been developed to reduce the calculation time. In this study, we show that ligand size may be used as an initial criterion for prioritizing ligands for docking. The ligands that are greatly different in size to that of the X-ray ligand, or to the binding site volume may be set to have low priority for screening or removed from the screening set, since these usually have low calculated binding affinity (K_i). On the other hand, ligands of optimum size within $\pm 1000 \text{ \AA}^3$ of the X-ray ligand or binding site volume may be assigned higher priority since they tend to bind with higher calculated affinity. In addition to clustering ligands into groups, adjusting the grid-box size would also help limit the number of ligands to be screened. The optimum grid-box allows one-third of the ligand to lie outside of the target binding site. Our new screening procedure, which takes into account the ligand size, the binding site volume, and the grid-box size, is easy to perform and could significantly reduce time and effort required for virtual drug screening.

Acknowledgment:

This investigation received financial support from the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Disease (TDR), National Center for Genetic Engineering and Biotechnology, Thailand (grant number BT-B-02-IG-GI-5101), and Thailand Research Fund (grant number DBG5180013) to Ekachai Jenwitheesuk. The author would like to thank Philip Shaw, Michal Guerquin, Siriphan Manochewa and people in the Biostatistics and Informatics laboratory for their critical reading and valuable comments.

References:

- [01] G. M. Morris and M. Lim-Wilby, *Methods Mol Biol.*, 443: 365 (2008) [PMID: 18446297]
- [02] J. Holyoake *et al.*, *Biophys J*, 91: 84 (2006) [PMID: 16980356]
- [03] K. P. Seiler *et al.*, *Nucleic Acids Res.*, 36: 351 (2008) [PMID: 17947324]
- [04] D. Dimitropoulos *et al.*, *Curr Protoc Bioinformatics*, 14: 14 (2006) [PMID: 18428761]
- [05] S. Goto *et al.*, *Nucleic Acids Res.*, 30: 402 (2002) [PMID: 11752349]
- [06] K. Arnold, *J Natl Cancer Inst.*, 95: 1192 (2003)[PMID: 12928341]
- [07] G. Schneider and H.J. Bohm, *Drug Discov. Today*, 7: 64 (2002) [PMID: 11790605]
- [08] P. Willett, *Drug Discov. Today*, 11: 1046 (2006) [PMID: 17129822]
- [09] M. A. Johnson and G.M. Maggiora, *John Wiley* (1990)
- [10] A. Talevi *et al.*, *J Enzyme Inhib Med Chem.*, 22: 253 (2007) [PMID: 17674806]
- [11] T. Kogej *et al.*, *J. Chem. Inf. Model*, 46: 1201 (2006) [PMID: 16711740]
- [12] G. W. Adamson and J. A. Bush, *J Chem Inf Comput Sci.*, 15: 15 (1975) [PMID: 1127038]
- [13] X. Chen and C. H. Reynolds, *J Chem Inf Comput Sci.*, 42: 1407 (2002)[PMID: 12444738]
- [14] M. von Grotthuss *et al.*, *Comb Chem High Throughput Screen*, 7: 757 (2004)[PMID: 15578937]
- [15] R. Guha *et al.*, *J Chem Inf Model*, 46: 99 (2006) [PMID: 16711717]
- [16] G. M. Morris *et al.*, *J Computational Chemistry*, 19: 1639 (1998)
- [17] H. M. Berman *et al.*, *Nucleic Acids Res.*, 28: 235 (2000)
- [18] N. Guex and M.C. Peitsch *Electrophoresis*, 18: 2714 (1997) [PMID: 9504803]
- [19] L. Kale *et al.*, *J Comp Phys.*, 151: 283 (1999)
- [20] J. Dundas *et al.*, *Nucleic Acids Res.*, 34: 116 (2006) [PMID: 16844972]
- [21] D. S. Goodsell *et al.*, *J Mol Recognit.*, 9: 1 (1996) [PMID: 8723313]

Edited by P. Kanguane

Citation: Wilantho *et al.*, *Bioinformatics* 3(5): 189-193 (2008)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Supplementary material



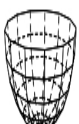

Binding site shape		PDB ID	
		Binding site volume	
		Small	Large
Simple		1OKE	NA
		1CET	NA
		1ZNY	1M4D
		1C3V	1LQU
		2FOM	1P44
		1ZAU	1EYE
		1K44	1F61
			2DEN
Complex	Irregular shape and varied depth	1R6A	1ENY
		1RQ7	
	L-shape with varied depth	1DF7	1MRN
			1N2B
			2C27
		1N8W	

Table 1: Shape and size of the protein binding sites is shown. The binding sites were divided into groups according to their shapes (simple and complex) and sizes (small; volume $<1200 \text{ \AA}^3$ and large; volume $\geq 1200 \text{ \AA}^3$).

PDB ID	Protein		Ligand		
	Name	Binding site volume(\AA^3)	Name	volume (\AA^3)	No. of active torsion
1C3V	Dihydrodipicolinate reductase	668	NA	NA	NA
2FOM	NS2B-NS3 protease	828	NA	NA	NA
1K44	Nucleoside diphosphate kinase	1012	NA	NA	NA
1F61	Isocitrate lyase	2735	NA	NA	NA
2DEN	Helicase	5071	NA	NA	NA
1R6A	Methyltransferase	305	RVP	1210	4
1RQ7	Filamenting temperature-sensitive mutant Z	594	GDP	1460	6
1ZNY	Guanylate kinase	664	GDP	1460	6
1DF7	Dihydrofolate reductase	752	MTX	1790	8
1ZAU	NAD ⁺ dependent DNA ligase	904	AMP	1240	4
1OKE	Envelope glycoprotein E	1044	BOG	1310	6
1CET	Lactate dehydrogenase	1074	CLQ	1410	8
1MRN	Thymidylate kinase	1268	T5A	2690	16
1M4D	Aminoglycoside 2'-N-acetyltransferase	1422	TOY	1680	6

1LQU	Mycobacterial oxidoreductase	1474	NDP	2400	13
1ENY	Enoyl-acyl carrier protein reductase	1657	NAD	2180	11
1P44	Enoyl reductase	1692	GEQ	1600	2
1N2B	Pantothenate synthetase	2008	APC	1570	8
1EYE	6-Hydroxymethyl-7,8-dihydropteroate synthase	2278	PMM	1040	3
2C27	Mycothioli synthase	3150	MA8	1540	6
1N8W	Malate synthase	5922	COA	2370	18

Table 2: Details of protein and ligand X-ray crystal structures used in this study.

PDB ID	Box center x, y, z axes	coordinate on x, y, z axes	No. of axis points on x, y, z axes
1C3V	134.796	25.327 22.087	40×50×40
2FOM	-4.451	-10.488 17.047	50×40×50
1K44	26.365	42.416 51.116	40×40×40
1F61	37.471	66.151 4.837	40×44×40
2DEN	-0.719	1.110 52.523	40×50×60
1R6A	16.498	-52.304 16.284	40×40×46
1RQ7	-8.274	37.324 7.065	40×40×40
1ZNY	24.782	8.849 30.726	40×40×40
1DF7	1.827	27.794 10.958	40×40×40
1ZAU	-10.553	43.872 68.997	40×40×40
1OKE	-11.491	80.375 45.517	60×40×40
1CET	35.592	12.727 18.358	40×40×40
1MRN	25.094	14.419 2.243	40×50×46
1M4D	20.818	29.061 13.435	40×40×40
1LQU	-4.075	-3.490 5.319	50×40×40
1ENY	-1.139	33.283 13.766	45×45×40
1P44	15.210	14.125 8.359	40×46×40
1N2B	34.878	35.004 40.786	40×40×40
1EYE	32.818	4.458 38.753	40×40×40
2C27	4.671	3.124 18.325	40×40×40
1N8W	15.581	32.585 77.100	40×50×40

Table 3: The coordinates of the grid-box center and the number of grid points on the x, y, and z axes in grid-boxes of minimum size is given. The distance between grid points was 0.375 Å.

PDB ID	Percentage of ligand atoms allowed outside of the box					
	5%	10%	15%	20%	25%	30%
1C3V	64.2	59.8	54.6	51.2	49.2	46.4
2FOM	59.2	40.0	23.4	13.2	8.6	5.6
1K44	73.4	60.6	45.8	30.6	18.8	11.4
1F61	88.4	76.6	65.2	50.4	38.2	27.2
2DEN	6.4	3.0	1.2	0.4	0.4	0.4
1R6A	50.8	36.8	24.2	16.2	11.8	8.4
1RQ7	52.0	39.0	25.6	16.8	9.0	5.8
1ZNY	78.4	63.2	51.8	39.2	27.6	18.6
1DF7	37.8	23.4	13.8	8.4	5.4	3.0
1ZAU	52.0	47.0	39.6	31.4	21.8	14.4

IOKE	12.8	7.6	3.6	1.8	1.0	0.8
1CET	61.4	47.6	34.8	23.2	15.0	8.4
1MRN	42.6	24.2	14.2	5.4	1.8	0.8
1M4D	38.0	29.0	19.8	13.6	9.2	5.2
1LQU	34.6	17.8	10.4	5.6	3.0	1.8
1ENY	40.2	22.6	11.8	5.6	2.2	1.2
1P44	41.4	31.6	22.4	14.0	9.6	6.0
1N2B	64.0	53.8	44.8	33.8	24.2	15.8
1EYE	95.6	89.2	78.8	64.2	49.4	39.6
2C27	42.8	29.0	20.4	14.0	6.2	3.6
1N8W	49.4	27.2	16.2	9.6	4.8	2.8
Average	51.6	39.4	29.6	21.3	15.1	10.8

Table 4: Percent of ligands rejected from the top500 ligand list docked with twenty-one protein binding sites and minimal grid-boxes (see Table 3). Six thresholds for rejection were tested, in which the percentage of ligand atoms outside of the grid-box was varied.

PDB ID	Ligand name	TC = 0.7		TC = 0.6		TC = 0.5	
		Top 500	Bottom 500	Top 500	Bottom 500	Top 500	Bottom 500
1R6A	RVP	1	0	19	0	47	0
1ZNY	GDP	0	0	1	1	2	14
1DF7	MTX	7	0	15	0	24	0
1ZAU	AMP	2	0	3	0	7	6
IOKE	BOG	0	0	0	0	0	0
1CET	CLQ	0	0	0	0	0	0
1MRN	T5A	1	0	11	2	23	16
1M4D	TOY	0	0	0	0	1	0
1LQU	NDP	8	0	13	4	21	4
1ENY	NAD	1	0	4	0	5	3
1P44	GEQ	0	0	1	0	28	0
1N2B	APC	0	0	7	8	13	9
1EYE	PMM	0	0	0	0	0	1
2C27	MA8	0	0	4	0	19	0
1N8W	COA	0	27	0	28	0	30
Total		20	27	78	43	190	83

Table 5: SMILES string similarity search of X-ray ligands in sixteen protein-ligand co-crystal X-ray structures against the top 500 and the bottom 500 docked ligands using Tanimoto coefficients (TC) of 0.5, 0.6, and 0.7, respectively.