Performing virtual screening by using multiple protein conformations

Aleksandra Olshanova
master student

The Laboratory of Molecular Modeling, Computer Aided Drug Design and Machine Learning

Bar-Ilan University, Israel

2022 – 2023

Contents

Introduction

Virtual screening (VS) is a powerful technique for identifying hit molecules as starting points for many industrial and academic drug discovery projects. Indeed, the number of methods and softwares which use different types of VS approaches is increasing at a rapid pace [1]. One type of them is structure - based virtual screening which plays a key role in the early stage of drug discovery. This process includes different methods to filter a chemical compound library using the structure of the molecular target of interest [2].

Available experimentally - determined 3D structures of the proteins are stored in the Protein Data Bank (PDB) archive, and the number of experimentally solved structures is increasing rapidly [3]. In order to obtain accurate docking results and consequently favorable VS results, it is important to have a protein structure from the PDB with a co-crystallized ligand which is chemically similar to target compounds available for screening. Of note, based on the chosen protein structure we can obtain various results for structure-based virtual screening. But even if we consider pairs of proteins with high sequence identity, they might have geometric differences over regions that are well-aligned in sequence [4]. At the same time identical protein structures might be crystallized with particular ligands which define different properties of the binding pocket. As a consequence this will lead to docking results which have significant differences for different conformations of the same protein.

The present project is focused on an assessment of docking performances across different protein structures of the acriflavine resistance B (AcrB) protein. AcrB is the inner membrane protein of the efflux complex of gram-negative bacteria and is responsible for the recognition and binding of compounds before their transportation out of the cell [5].

Target protein for structure-based virtual screening

The multidrug resistance in bacteria has become one of the major threats in global health [6]. Bacteria have developed a mechanism to evade the attack of most commercially available drugs. In gram-negative bacteria such as *Escherichia coli*, multidrug efflux systems are a major mechanism that confers intrinsic and increased drug resistance to a broad spectrum of antimicrobials. AcrB is the inner membrane protein of the efflux complex and is responsible for the recognition and binding of compounds before their transportation out of the cell [5]. In this project the docking calculations were performed on the so-called “hydrophobic trap” pocket of the AcrB protein structure of the *Escherichia coli* bacteria (strain K12).

The hydrophilic substrate translocation channel and hydrophobic trap are two key features of the AcrB protein that enable it to bind and export substrates efficiently. The hydrophilic channel is lined with polar amino acid residues that interact with the polar or charged regions of the substrate molecules, facilitating their translocation across the cell membrane. The hydrophobic trap, on the other hand, is located within the protein's interior and is composed of non-polar amino acid residues. This trap is designed to capture hydrophobic or amphipathic substrates, preventing their backflow into the cell. Inhibitors of the AcrB protein often target these two regions to prevent substrate binding and export.

The importance of the trap was first recognized by Nakashima et. al. (2013) when they published the first co-crystal structure of the inhibitor, pyridopyrimidine EPI D13-9001 (also known in the literature as ABI-PP), bound to both AcrB from E. coli and MexB from P. aeruginosa. A key amino acid in the hydrophobic trap is Phe 178, which plays an important role in inhibitor binding through π-π stacking interactions with aromatic substrates [7]. The discovery of this hydrophobic trap is considered an important advance, contributing to the effective development of inhibitors using structure-guided design. The contribution of various residues to binding is summarized in **Figure 1**. Residues F178, I277, V612, and F615 contributed most to the stabilization of almost all known inhibitors.

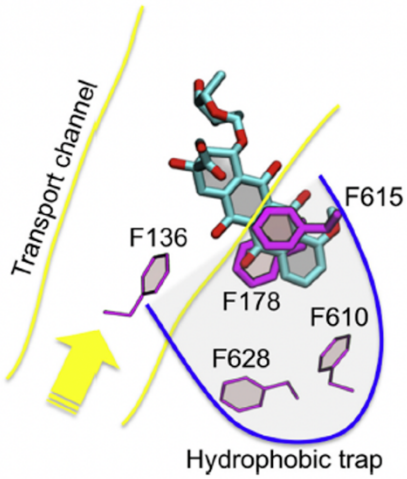


Figure 1. Doxorubicin bound to the AcrB in hydrothobic trap [8].

Frequency of contribution to the binding free energy from different residues belonging to the key regions and from additional residues is presented in **Figure 2**.

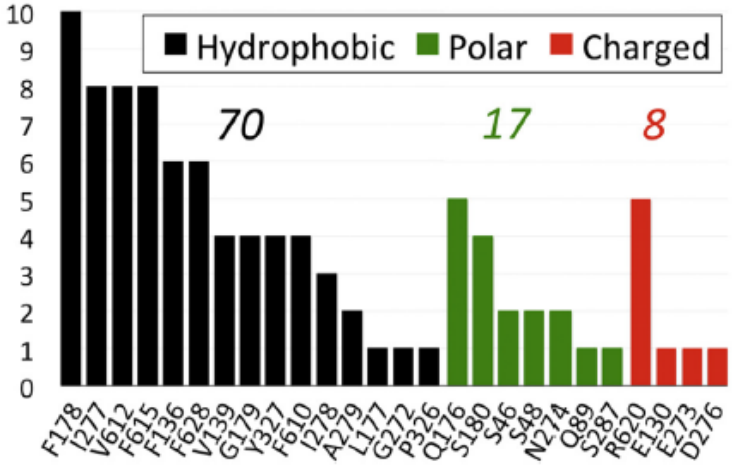


Figure 2. Frequency of contribution to binding free energy of substrates by hydrophobic (black bars), polar (green) and charged (red) residues. The sum over all frequencies is reported above each part of the histogram [9].

Materials & Methods

ROC curve model validation

The “receiver operating characteristic” (ROC) curve method is a well-recognized metric used as an objective way to evaluate the ability of a given test to discriminate between two populations. This technique facilitates decision-making in various fields including medicinal chemistry. For instance, when virtual screening is used to boost the drug discovery process in pharmaceutical research, taking the right decision upon selecting or discarding a molecule prior to in vitro evaluation is of paramount importance. Characterizing both the ability of a virtual screening workflow to select active molecules and the ability to discard inactive ones, the ROC curve approach is a well - suited metric for evaluating the confidence with which this decision could be taken [10].

The assessment of a computer test for virtual screening by the ROC curves method requires four steps:

Step 1. Definition of active and inactive compounds.

Depending on the experimental method that will be further used to evaluate the selected molecules, it is necessary to define an appropriate cutoff between active compounds and those considered as inactive for the target.

Step 2. Selection of a set of mixed compounds.

The next step requires the selection of a set of molecules (containing both active and decoy, yet presumed to be inactive compounds) of relevance to the target under investigation. Ideally, it should contain structurally diverse compounds of known activity in order to cover the chemical space as much as possible. Also, decoy compounds molecules should not be randomly picked but should rather have a chemical structure similar to the structures of the chosen actives [11].

Step 3. Virtual screening of the set.

Docking and scoring all compounds (obtained set) against the target binding site using docking programs.

Step 4. ROC curve analysis.

Finally, the ROC curve method is applied to assess the performance of the docking process. **Figure 3** illustrates the overall technique for theoretical distributions of active and inactive compounds. The concept of confusion matrices in the context of drug design is used as a tool to comprehend the ROC curve method better as it allows quick calculation of sensitivity and

specificity from a comparison between in vitro (active/inactive) and in silico (selected/discarded) classifications [12].

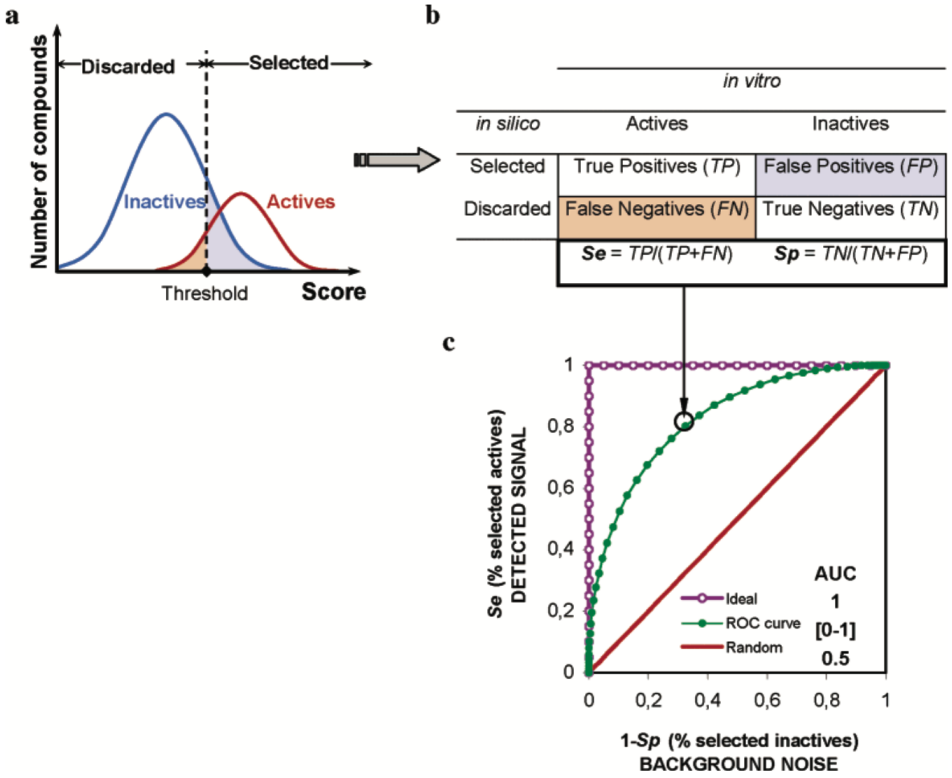


Figure 3. The overall technique for theoretical distributions of actives and inactive compounds. For a given selection scoring threshold (panel a), the classification of all compounds is reported in a "confusion matrix" (panel b). For all possible score thresholds, the evolution of the deduced sensitivity (Se) and specificity (Sp) is reported on a ROC graph, Se as a function of 1 – Sp (panel c) [10].

The area under the curve (AUC) is a practical way of measuring the overall performance of the tests. If the AUC is close to 0.5 (the value of a random test), the test is said to be poor; the highest possible AUC is 1, corresponding to an ideal case. In general, the greater the AUC, the more effective the virtual screening workflow is in discriminating active from inactive compounds.

Finally, the ROC curve analysis helps answer the question: Considering current knowledge, how good is my model at selecting the known active molecules and discarding the inactive ones compared to another model?

Once the theoretical aspect has been depicted, the ROC curve method will be illustrated by the construction of a virtual screening workflow focused on the acriflavine resistance B (AcrB) protein inhibitors.

Preparation of the Protein Files

Experimentally-determined 3D structures of AcrB protein (*Escherichia coli* K12) were taken from the Protein Data Bank (PDB) including 5ENS [13], 3W9H [7], 5ENQ [13], 5ENR [13], 3AOD [14], 5ENT [13], 7B8R [15], 7B8T [15]. The receptor preparation for all 8 protein structures was carried out with Schrodinger's Protein Preparation Wizard. First, all water molecules were removed, as well as ligand atoms, and those ions that do not belong to the active site of the receptor from the corresponding PDB file. Next, hydrogen atoms were added, and amino acid side-chain formal charges (in accordance with the protonation state at the physiological pH) were assigned. This was followed by a local minimization to relieve potential bad contacts. The minimization was performed in the presence of restraints to maintain the protein conformation very close to that observed in the experimental model.

Preparation of the Ligand Files

Active compounds for docking were provided by ChEMBL database of bioactive molecules with drug-like properties [16], [17]. Decoy compounds were generated by the DUD-E database. The ligand structures were converted to the SMILES notation with the aim of avoiding any bias from the bioactive conformation in conformer calculations. Prior to docking, hits were prepared by the LigPrep program as implemented in Maestro (Schrödinger, USA) at pH = 7 ± 0.2 with the OPLS 2005 force field, including tautomeric variations. Ligand coordinate files were extracted from the corresponding PDB files and used as reference structures for root mean square deviation (RMSD) calculations to validate the docking tool.

Docking Algorithms

Docking was performed using Standard-Precision (SP) Glide with default settings. Number of poses for each ligand was set to 10.

Clustering and ROC curve

Clustering analysis and a "receiver operating characteristic" curve (ROC) were performed by Python.

Results & Discussion

Identifying potential active sites with SiteMap

Top - ranked potential receptor binding sites were predicted by running Schrodinger's SiteMap for the above-mentioned eight protein structures following the removal of their ligands. Here, a computational method helped to identify the location of possible binding sites for each protein structure, and then by protein structure alignment determine which sites are similar to each other. Overall, 11 potential binding sites were generated by SiteMap calculations. The potential regions suitable for binding are illustrated in **Figure 4**.

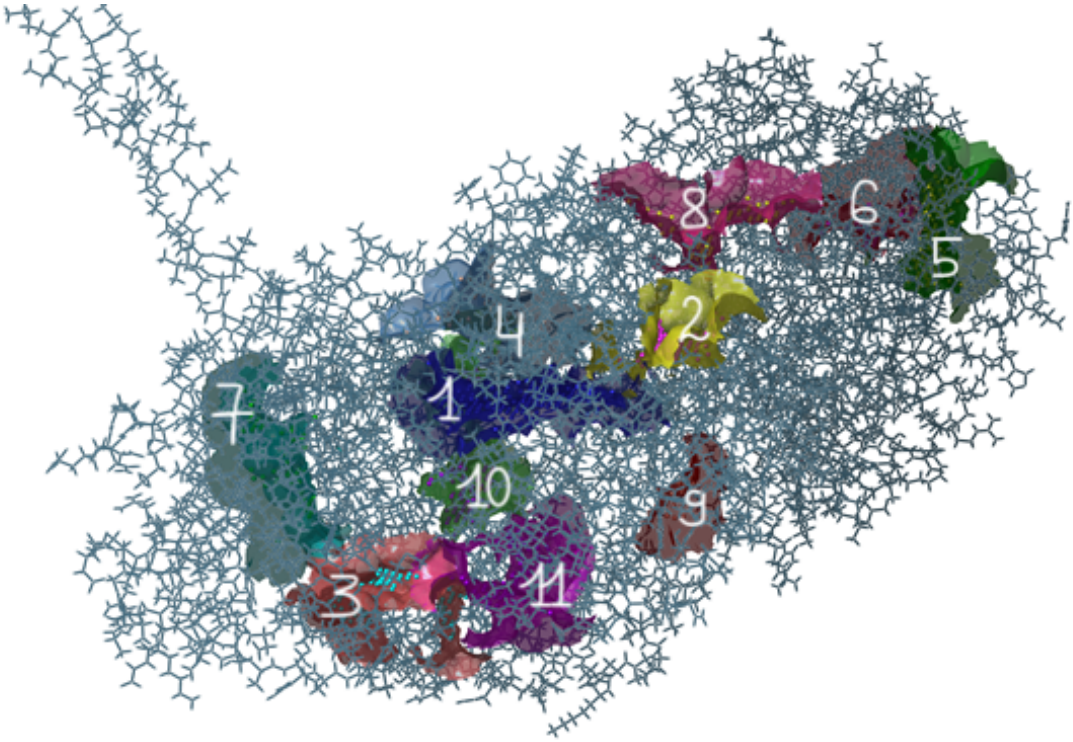


Figure 4. Potentially druggable sites surfaces found over eight protein structures (presented using the 3AOD protein structure).

To evaluate the primary binding site, it is recommended to look for a SiteScore and Dscore > 1, a balance > 0.3 and a volume > 225. Moreover, the most important metric generated by SiteMap is an overall SiteScore, which has proven to be effective at identifying known binding sites in co-crystallized complexes. According to the obtained results, the most promising pocket is the first one, as it is characterized by the highest SiteScore mean value among all structures (**Table 1**). Also, it has mean values for other calculated properties in the desired ranges: Dscore > 1, volume > 225, balance > 0.3. Indeed, the first binding site is the distal binding pocket, which is named as hydrophobic trap, and has a high consideration for the design of small molecule inhibitors that target AcrB [18].

Table 1. Mean values of properties generated by SiteMap in descending order of SiteScore values.

Binding site	Protein structures with found site	SiteScore	Dscore	volume	balance
1	5ENS, 3W9H, 5ENQ, 5ENR, 3AOD, 5ENT,7B8R, 7B8T	1.15 ± 0.05	1.23 ± 0.07	603.22 ± 136.34	4.03 ± 2.87
5	3W9H	1.10	1.03	270.28	0.37
3	3W9H, 5ENR, 3AOD, 5ENT	1.04 ± 0.01	0.98 ± 0.07	330.57 ± 70.73	0.26 ± 0.13
8	3W9H	1.01	1.09	214.03	7.41
10	5ENR, 5ENT,7B8R, 7B8T	0.91 ± 0.10	0.87 ± 0.09	336.23 ± 86.78	0.13 ± 0.03
11	5ENQ, 7B8T	0.90 ± 0.03	0.84 ± 0.04	246.10 ± 4.61	0.49 ± 0.35
7	5ENS, 5ENQ, 5ENR, 7B8T	0.86 ± 0.08	0.84 ± 0.11	258.11 ± 55.13	0.52 ± 0.31
2	3W9H, 5ENQ, 3AOD, 7B8R	0.80 ± 0.19	0.76 ± 0.21	174.76 ± 62.78	0.26 ± 0.18
6	3AOD	0.79	0.60	162.58	0.33
4	5ENS, 5ENT, 7B8R, 7B8T	0.76 ± 0.08	0.70 ± 0.15	124.92 ± 32.66	0.26 ± 0.19
9	7B8T	0.58	0.52	90.55	0.29

After identifying all possible binding sites over all protein structures, SiteMap was run in a different mode which allows to evaluate the region of the receptor that is within a specified distance (6 Å) of co - crystallized ligand. As all protein structures contain ligands in the first binding site, it was possible to align the obtained potential surfaces and calculate mean values of key properties for them. The alignment of the first site's surfaces over eight protein structures is illustrated in **Figure 5**.

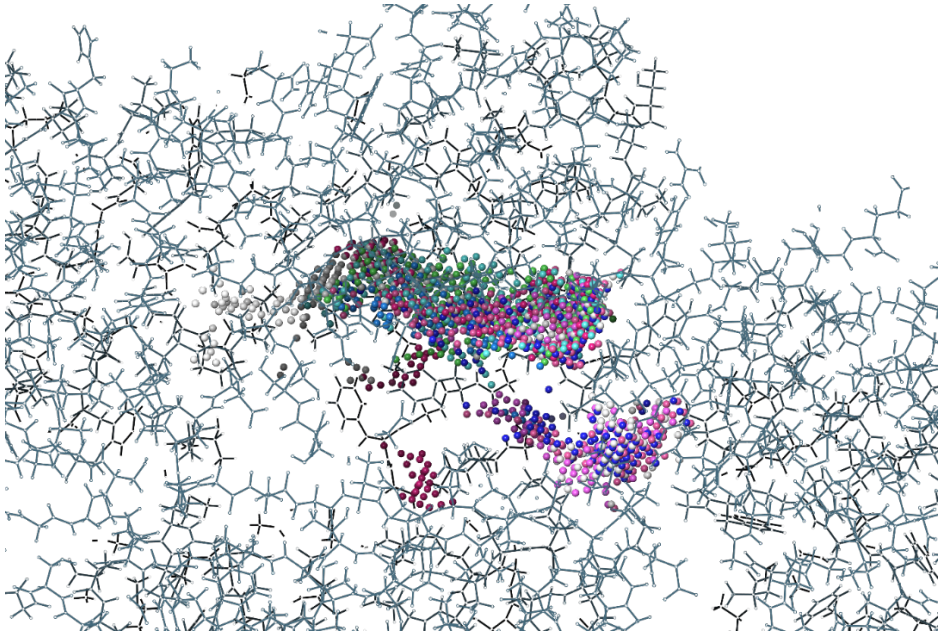


Figure 5. Aligned first binding site surfaces over eight protein structures (presented using the 3AOD protein structure).

Calculated average values of key properties (**Table 2**) correspond with the results from the first analysis where all top - ranked potential receptor binding sites were predicted.

Particularly, mean values of SideScore, Dscore, volume and balance for the first binding site are almost equal in the two analyses (the SiteScore mean value differ by 0.02). Eventually, SiteMap has suggested that the first binding site is the most desirable among all 11 potential binding sites, and will be used later to perform docking.

Table 2. Mean values of key properties generated by SiteMap for the first binding site over all protein structures.

	SiteScore	Dscore	volume	balance
count	8	8	8	8
mean	1.13	1.19	580.27	3.23
std	0.06	0.08	174.54	2.37
min	1.06	1.10	383.13	0.86
max	1.21	1.31	939.82	7.29

When performing target analysis on a ligand - bound structure, it is a best practice to run SiteMap in both modes (searching for top-ranked sites and a single site), and then look for similar results for the desired site.

Performing hierarchical clustering

Hierarchical clustering was used to select a small and representative subset from within the set of eight structures of AcrB proteins (*Escherichia coli* K12). After identifying patterns of the degree of similarity between structures, similar protein structures were merged into a single cluster to decrease the number of docking experiments.

First, all protein structures were aligned based on their first binding site, and then root-mean-square deviation (RMSD) values were calculated. “Align binding sites” panel first runs a global alignment and then automatically generates the list of atoms to use in pairwise alignment from the residues, which were selected based on the distance from the ligand (within 5.0 Å from the ligand). Then in-place RMSD values for each pair of structures were calculated ignoring atom pairs greater than 5.0 Å apart. The RMSD matrix between all proteins is presented in **Table 3**, and then was used to perform clustering.

Table 3. RMSD values (in Å) of aligned protein structures based on their first binding site.

RMSD	5ENS	3W9H	5ENQ	5ENR	3AOD	5ENT	7B8R	7B8T
5ENS	0							
3W9H	0.910	0						
5ENQ	0.464	0.946	0					
5ENR	0.401	0.907	0.268	0				
3AOD	1.055	0.834	1.140	1.112	0			
5ENT	0.337	1.099	0.531	0.451	1.212	0		
7B8R	0.402	1.078	0.563	0.478	1.220	0.398	0	
7B8T	0.401	0.984	0.605	0.554	1.163	0.393	0.460	0

Then the hierarchical clustering technique was performed, and predicted clusters were visualized using a dendrogram (**Figure 6**). In this project I implemented procedures to perform a Ward´s linkage clustering, which is a variance-based method which attempts to generate clusters to minimize the within-cluster variance.

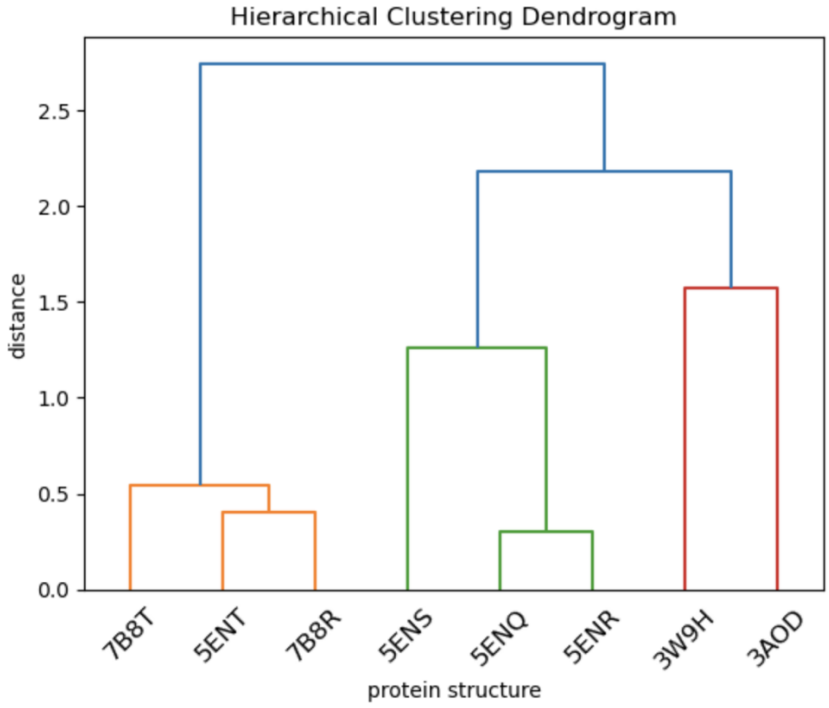


Figure 6. Hierarchical clustering dendrogram.

According to the dendrogram three individual clusters were identified, and the representative protein structure for each group was defined (**Table 4**).

Table 4. Grouped protein structures based on their binding site's similarity.

№ of cluster	Protein structures	Structure for docking
1	7B8T, 5ENT, 7B8R	5ENT
2	5ENS, 5ENQ, 5ENR	5ENR
3	3W9H, 3AOD	3W9H

Hierarchical clustering technique was chosen, because it has some benefits over other types of clustering as not having to pre-specify the number of clusters and the fact that it can produce a nice hierarchical illustration of the clusters.

Docking with Glide

For predicting protein-ligand binding modes and ranking ligands via high-throughput virtual screening, the Glide (SP) docking program was used. The stages involved in docking are:

1) Protein preparation with Protein Preparation Wizard
2) Ligand preparation with LigPrep
3) Grid Generation
4) Docking the self-ligand
5) Docking the screening compounds

Self-docking study

First of all, a self-docking study was undertaken for one representative structure from each of the three clusters. Crucial step in docking is the generation of a receptor grid which determines where, on the protein surface, the ligands will be docked.

The three ligands presented in **Figure 7** ([7], [13], [19]) were converted into their SMILES representation, prepared with LigPrep, and then subjected to docking calculations to their respective protein structure grids.

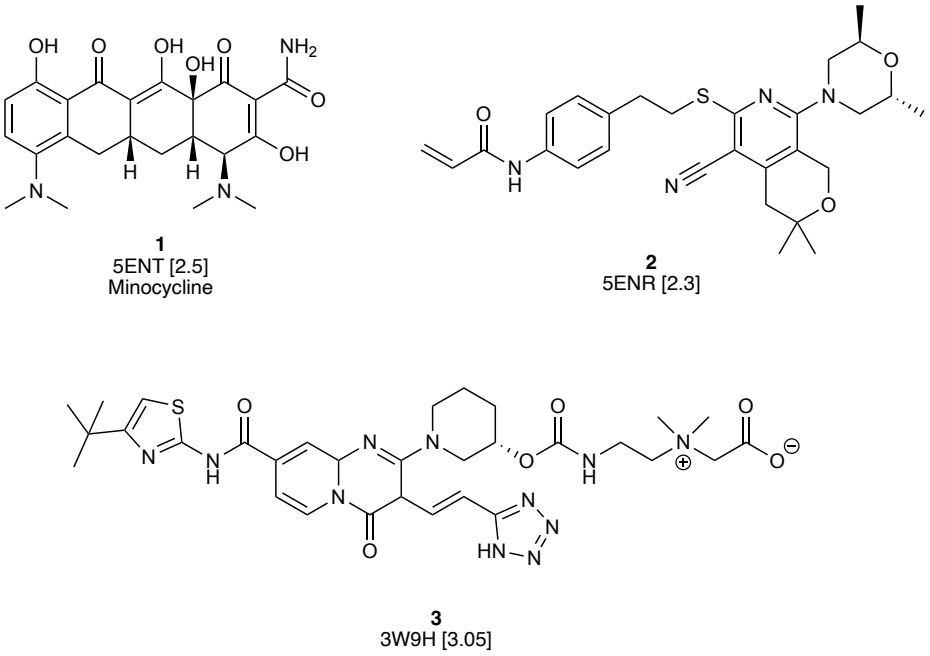


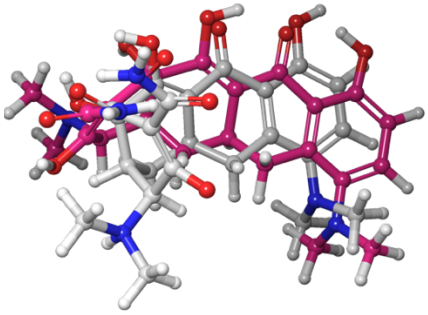
Figure 7. Selected ligands, with the PDB code and resolution (in Å, within brackets) of the crystalline complexes.

All obtained conformations of the ligands were compared with the crystallized conformations from the PDB protein crystal structures to evaluate the ability of the Glide program to reproduce the experimentally determined protein – ligand crystal structures. The root mean square deviation values were calculated for each pair of ligands (**Table 5**).

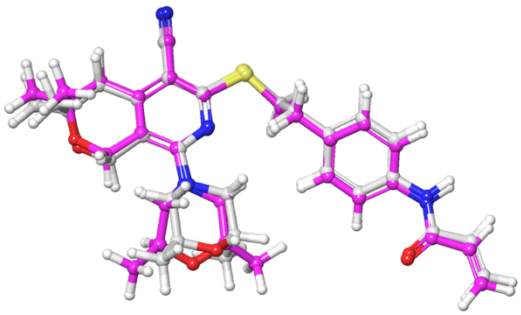
Table 5. RMSD values between the lowest energy pose and the crystallized conformation from PDB protein crystal structures, and their corresponding glide emodel score.

№ of cluster	Glide emodel (kcal / mol)	RMSD (Å)
1	-53.904	7.165
2	-107.608	0.711
3	-124.525	2.722

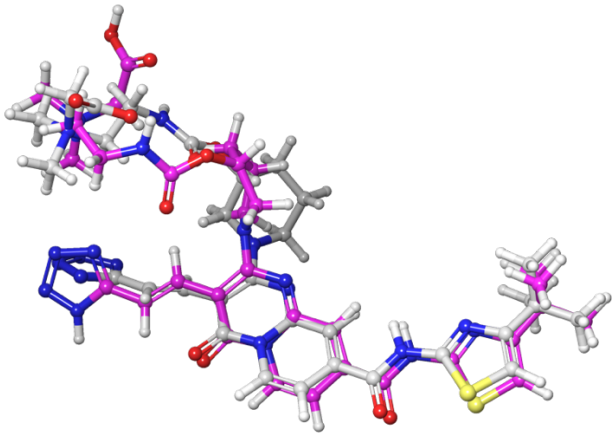
In fact, the root mean square deviation value was quite low (RMSD ≤ 1.0 Å) for the structure representing the second cluster (5ENR protein structure), which means that Glide is capable of reproducing the experimentally determined protein - ligand crystal structure for this protein. However, for the other proteins the same result was not obtained. According to the high RMSD values Glide failed in reproducing protein-ligand crystal structures for 5ENT and 3W9H proteins, which is illustrated in **Figure 8**.



5ENT



5ENR



3W9H

Figure 8. In-place superposition of the lowest energy pose of the ligand after self-docking (grey color) and crystallized conformation from PDB protein crystal structures (pink color) for representative structures from each cluster.

The ligand from the 3W9H protein structure has many rotatable bonds, which might be a reason for having a high RMSD value. However, the ligand from 5ENT is rigid, but the RMSD value is still high. Additional calculations were performed for this ligand, where the ligand was extracted from crystal protein structure and then subjected to rigid docking calculation using its own bound conformation. The obtained RMSD value was 0.98 (Å), which is good for this structure. Thus, the reason for the high RMSD value for ligands generated from their SMILES representation could be attributed to LigPrep ligand preparation procedure, which somehow changes the ligand conformation. Alternatively, the sampling procedure employed by Glide was insufficient in this case.

Docking the screening library

All prepared active compounds from the ChEMBL database and the decoy compounds generated by the DUD-E database were docked into the docking grids produced for the three structures. All together 63 active compounds from the ChEMBL database and 2653 generated decoy compounds were docked. The Box plot method was used to graphically show the distribution of numerical Glide gscore values and to visualize differences in docking scores between the three representative structures (**Figure 9)**.

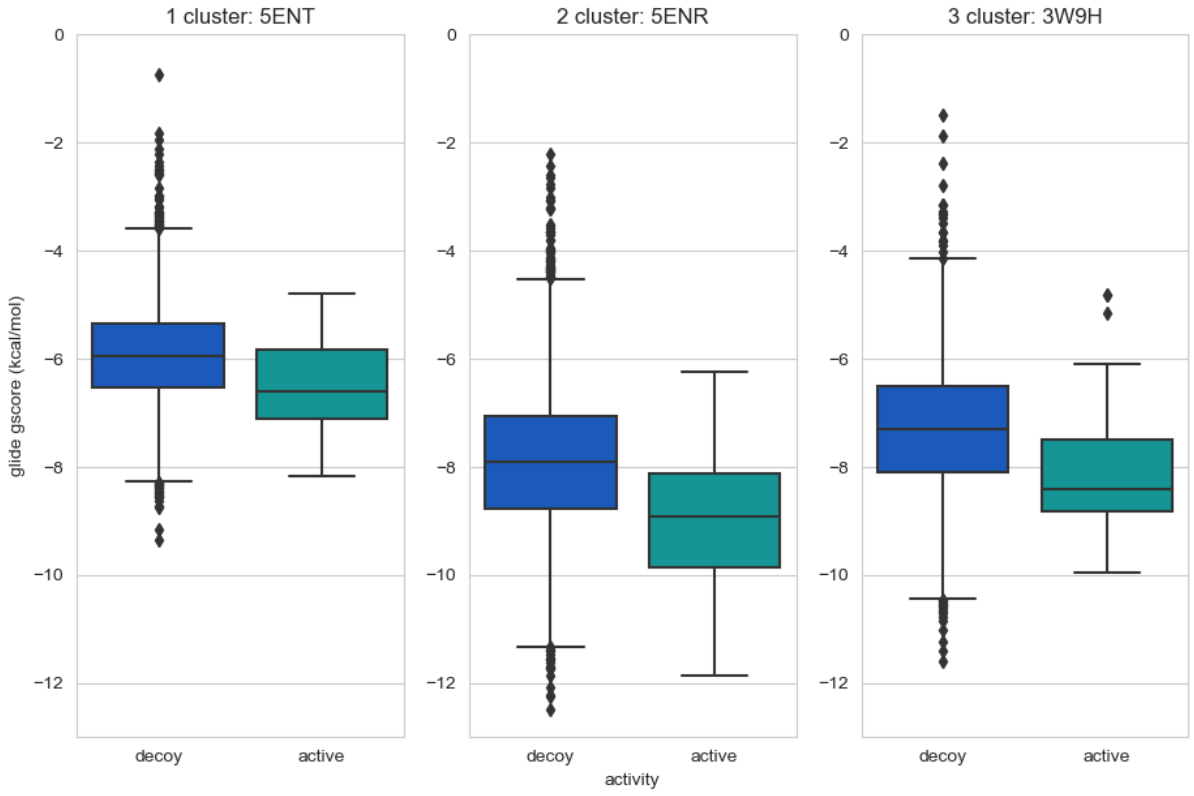


Figure 9. The distribution of Glide gscore data for the representative proteins from each cluster: 5ENT, 5ENR, 3W9H.

The whiskers of the box plot extend from the box to the minimum and maximum values of the data that are not considered outliers. Indeed, obtained data of glide gscores for active compounds does not have outliers for each representative structure from the first and second cluster, which means that Glide is capable of recognizing active compounds, and their glide scores have similar values. The representative structure 3W9H from the third cluster has some outliers for active compound glide score, which might result from a low quality of the grid box which was generated from the protein structure and the co-crystallized ligand which has many rotatable bonds.

Outliers, or extreme values that fall outside the range of the whiskers, are represented by individual points beyond the whiskers. Such outliers are observed for the data of Glide gscores for decoy compounds for each representative structure from each of the three clusters. This is an expected behavior for inactive compounds, which means that Glide is capable of identifying decoy compounds.

Model validation

In this project, the "receiver operating characteristic" curve (ROC) was used to compare the performance of VS process in attributing better scores to active than inactive ligands. After performing docking for each cluster, the corresponding ROC curve was produced. ROC curves for each of the representative structures are illustrated in **Figure 10**.

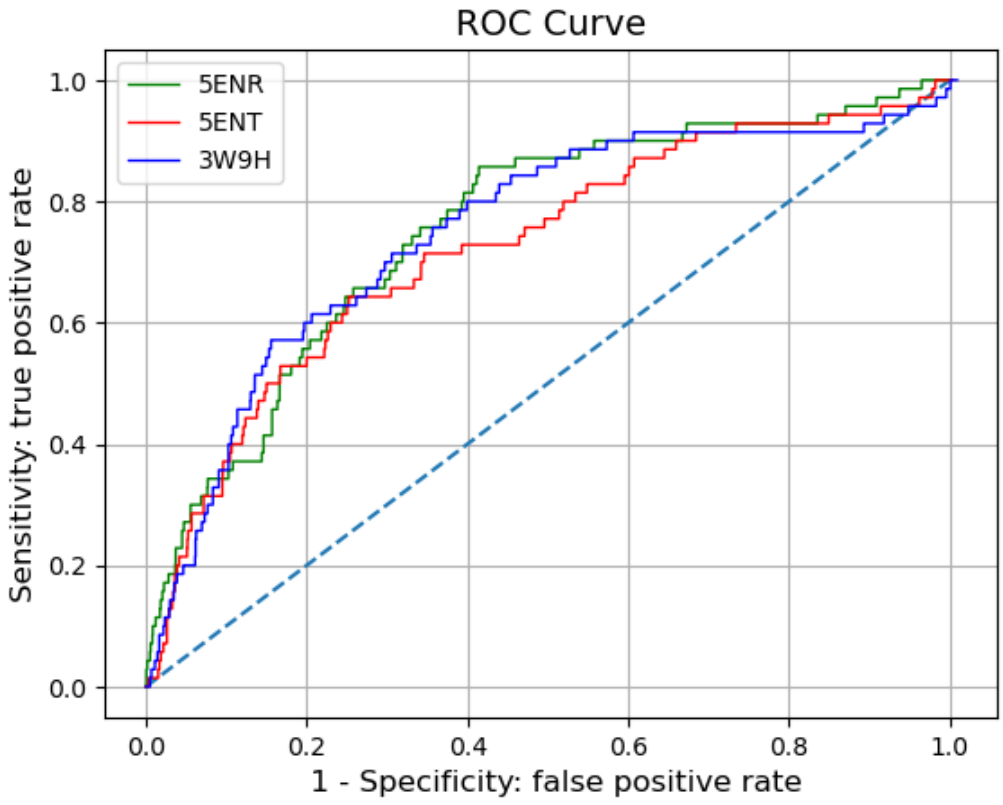


Figure 10. ROC curves for each representative structure.

The area under the curve (AUC) allows to evaluate the probability that a randomly chosen known active will rank higher than a randomly chosen decoy. According to the calculated AUC values for each cluster (**Table 6**) the probability over all clusters is around 0.73-0.76.

Table 6. Calculated AUC values for each cluster.

Cluster	Representative protein structure	AUC
1	5ENT	0.7269
2	5ENR	0.7554
3	3W9H	0.7581

Conclusions

The main purpose of this project was the estimation of how the Glide docking program is capable of performing docking calculations for different protein structures of the same protein (AcrB). According to the obtained results, the area under the curve (AUC) is pretty much the same for all three representative structures.

ROC curves analysis shows that, in the present case, different protein structures do not influence significantly Glide's docking performance. However, from the obtained results, the representative structure 5ENR of the second cluster has the best overall characteristics. First, a self-docking study was successful for this protein structure (the root mean square deviation value for the lowest energy pose was quite low (RMSD $\leq$ 1.0 Å). Indeed, Glide was able to reproduce the experimentally determined protein - ligand crystal structure. Second, the AUC value was slightly higher compared with that for the first structure, which means that Glide can slightly better evaluate the differences between active and decoy compounds using this protein structure. Thus, for the purpose of a large-scale virtual screening campaign, this would be the protein structure of choice.

References

[1] Lavecchia A., Di Giovanni C. Virtual screening strategies in drug discovery: A critical review. // Current medicinal chemistry. **2013**. V.20. №23. P. 2839–2860.

[2] Kitchen D. B. et. al. Docking and scoring in virtual screening for drug discovery: Methods and applications. // Nature Reviews Drug Discovery. **2004**. V. 3. №11. P. 935–949.

[3] Burley S. K. et. al. Protein data bank: A comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. // Biomolecules. **2022**. V. 12. №10. P. 1425.

[4] Kosloff M., Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. // Proteins: structure, function and genetics. **2008**. V. 71. №2. P. 891–902.

[5] Rajapaksha P. et. al. Probing the dynamics of AcrB through disulfide bond formation. // ACS Omega. **2020**. V. 5. №34. P. 21844–21852.

[6] Vivas R. et. al. Multidrug-resistant bacteria and alternative methods to control them: An overview // Microbial Drug Resistance. **2019**. V. 25. №6. P. 890–908.

[7] Nakashima R. et al. Structural basis for the inhibition of bacterial multidrug exporters. // Nature. **2013**. V. 500. №7460. P. 102–106.

[8] Vargiu A. V. et al. Computer simulations of the activity of RND efflux pumps. // Research in Microbiology. **2018**. V. 169. №7–8. P. 384–392.

[9] Vargiu A. V., Nikaido H. Multidrug binding properties of the AcrB efflux pump characterized by molecular dynamics simulations. // Proceedings of the National Academy of Sciences (PNAS). **2012**. V. 109. №50. P. 20637–20642.

[10] Triballeau N. et al. Virtual screening workflow development guided by the 'receiver operating characteristic' curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. // Journal of Medicinal Chemistry. **2005**. V. 48. №7. P. 2534–2547.

[11] Verdonk M. L. et al. Virtual screening using protein-ligand docking: Avoiding artificial enrichment. // Journal of chemical information and computer sciences. **2004**. V. 44. №3, P. 793–806.

[12] Manallack D. T. et al. Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. // Journal of chemical information and computer sciences. **2002**. V. 42. №5. P. 1256–1262.

[13] Sjuts H. et al. Molecular basis for inhibition of AcrB multidrug efflux pump by novel and powerful pyranopyridine derivatives. // Proceedings of the National Academy of Sciences (PNAS). **2016**. V. 113. №13. P. 3509–3514.

[14] Nakashima R. et al. Structures of the multidrug exporter AcrB reveal a proximal multisite drug-binding pocket. // Nature. **2011**. V. 480. №7378. P. 565–569.

[15] Ornik-Cha A. et al. Structural and functional analysis of the promiscuous AcrB and AdeB efflux pumps suggests different drug binding mechanisms. // Nature Communications. **2021**. V. 12. №1.

[16] Jin C. et al. Design, synthesis and evaluation of a series of 5-methoxy-2,3-naphthalimide derivatives as AcrB inhibitors for the reversal of bacterial resistance. // Bioorganic & Medicinal Chemistry Letters. **2019**. V. 29. №7. P. 882–889.
[17] Wang Y. et al. Design and structural optimization of novel 2H-benzo[h]chromene derivatives that target AcrB and reverse bacterial multidrug resistance. // European Journal of Medicinal Chemistry. **2021**. V. 213.
[18] Alenazy R. Drug efflux pump inhibitors: A promising approach to counter multidrug resistance in gram-negative pathogens by targeting AcrB protein from AcrAB-TolC multidrug efflux pump from Escherichia coli. // Biology. **2022**. V. 11. №9.
[19] Takatsuka Y. et al. Mechanism of recognition of compounds of diverse structures by the multidrug efflux pump AcrB of Escherichia coli. // Proceedings of the National Academy of Sciences (PNAS). **2010**. V. 107. №15. P. 6559–6565.