

**Name: Konstantinos Giatras**  
**Course: Molecular Modeling of Biomolecules**  
**Final Project**  
**Date: 20/01/2023**

**1. Perform Protein Preparation on the cluster representative of your PTEN simulation. Explain the procedure and the steps. (1 point)**

We load the ionized.psf file along with the 1d5r\_wb\_prod.dcd trajectory file that we have previously created in VMD, but only frame 70 (our cluster representative of our PTEN production dynamics simulation), which we found from the trajectory analysis (clustering) to be the centroid of the cluster with the smallest average RMSD. We then save the coordinates as a new pdb file, which we are going to be using in Maestro.

The Maestro program is a comprehensive software suite for computer-aided drug design and molecular modeling. It includes a variety of tools for tasks such as protein structure preparation, molecular dynamics simulation, virtual screening, and ligand-receptor docking. It provides a user-friendly interface for accessing these tools, and also includes a variety of visualization tools for analyzing and interpreting results.<sup>[1]</sup>

Protein preparation is a crucial step in the drug discovery process, as it involves the preparation of a high-quality 3D structure of a target protein that will be used for molecular simulations and virtual screening. The goal of protein preparation is to produce an accurate and biologically relevant model of the protein that accurately represents its active conformation, taking into account its overall shape, electrostatic potential, and other important properties. This information is then used to guide the design of drugs that bind to the protein and modulate its activity.<sup>[1]</sup>

We load our pdb file into Maestro and begin the Protein Preparation. In this step (Protein Preparation Workflow), we choose the appropriate options to perform the following: we delete the waters if we haven't done so beforehand, we assign bond orders using the CCD database, we fill in missing side chains, we replace the hydrogens, we optimize h-bond assignments using PROPKA, we create zero-order bonds to metals and disulfide bonds, we generate taut states (with Epik) for pH: 7.4 +/- 2.0 (max states to process automatically: 1) and we also minimize by converging heavy atoms to RMSD: 0.30 Å. We click run and wait for the job to complete, after which the prepared protein appears in our Maestro workspace. We can always monitor the progress of each job in Maestro's job tab.

**2. Perform Binding Site detection on your protein using SiteMap from Maestro. Explain briefly the procedure (from the manual) and the results. Provide pictures of the results. (1 point)**

After running Protein Preparation, we perform Binding Site detection on our protein using SiteMap, which is a computational method for identifying binding pockets on a protein structure.

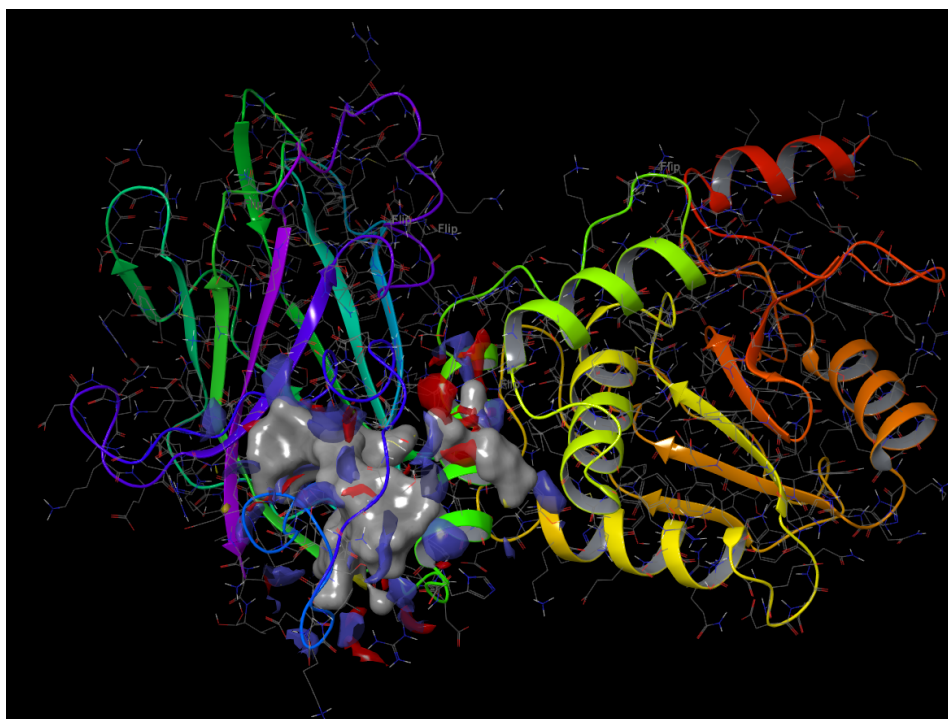
It uses a combination of geometric and physico-chemical criteria to identify potential binding sites on a protein. The general process of SiteMap is as follows:

1. The protein structure is first processed to remove any water molecules, ligands, and cofactors that may be present.
2. The protein is then divided into a grid of voxels (3D pixels), and the properties of each voxel are calculated based on the atoms present.
3. The voxels are then grouped into clusters based on their properties. Clusters with similar properties are merged together to form larger clusters.
4. The resulting clusters are then screened to identify those that are likely to represent binding pockets. Clusters that meet certain geometric and physical-chemical criteria (e.g. size, shape, electrostatic potential, etc.) are considered as potential binding pockets.
5. Finally, the identified binding pockets are refined using molecular dynamics simulations or other methods to optimize their shape and electrostatic properties.<sup>[2]</sup>

In Maestro's tasks, we search for SiteMap and click on it, which opens a new window. After making sure we have selected our protein, we run the application with its default options. Its output appears in our workspace after completion. For our protein, it identifies 5 binding sites:

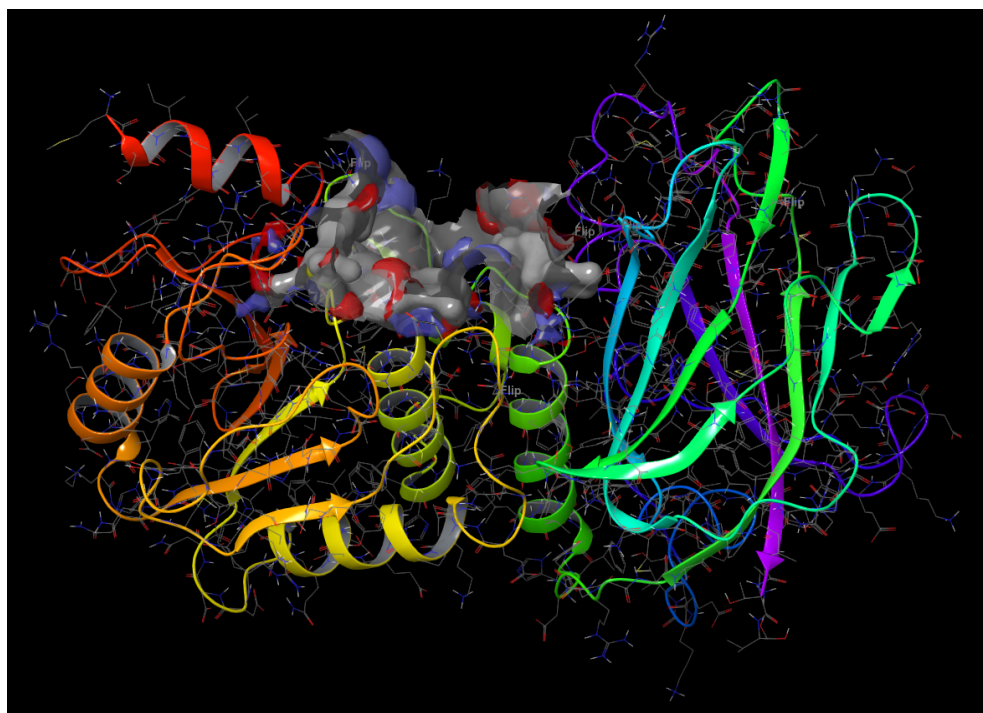
| Row | In                               | Title                   | Stars | volume  | size | Dscore | SiteScore | Entry ID |
|-----|----------------------------------|-------------------------|-------|---------|------|--------|-----------|----------|
| 1   | <input type="radio"/>            | 1d5r_frame70            | ☆☆☆   |         |      |        |           | 1        |
| 1   | <input checked="" type="radio"/> | proteinprep_2-out1 (1)  |       |         |      |        |           |          |
| 2   | <input checked="" type="radio"/> | 1d5r_frame70 - prepared | ☆☆☆   |         |      |        |           | 2        |
| 1   | <input checked="" type="radio"/> | sitemap_1_out1 (6)      |       |         |      |        |           |          |
| 3   | <input checked="" type="radio"/> | sitemap_1_site_2        | ☆☆☆   | 284.690 | 97   | 1.058  | 1.016     | 3        |
| 4   | <input type="radio"/>            | sitemap_1_site_1        | ☆☆☆   | 493.234 | 206  | 1.046  | 1.003     | 4        |
| 5   | <input type="radio"/>            | sitemap_1_site_4        | ☆☆☆   | 147.147 | 70   | 0.838  | 0.907     | 5        |
| 6   | <input type="radio"/>            | sitemap_1_site_5        | ☆☆☆   | 176.302 | 65   | 0.839  | 0.881     | 6        |
| 7   | <input type="radio"/>            | sitemap_1_site_3        | ☆☆☆   | 124.166 | 67   | 0.868  | 0.852     | 7        |
| 8   | <input type="radio"/>            | sitemap_1_protein       | ☆☆☆   |         |      |        |           | 8        |

Site 1:



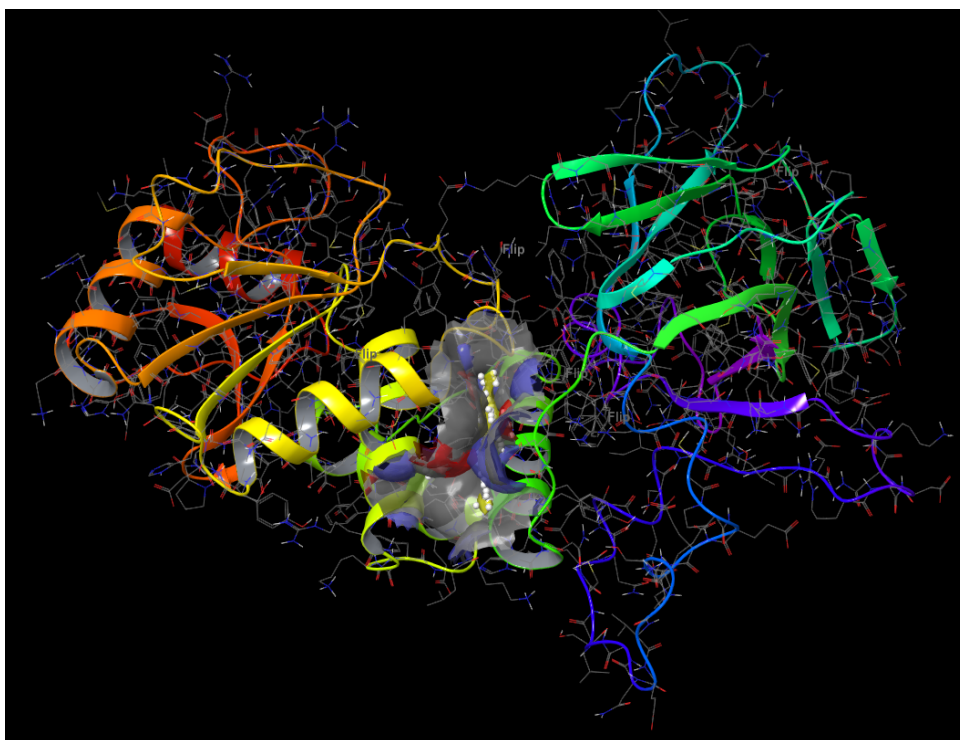
Involved residues (after expanding 3 Å onto the protein for each identified binding site):  
LEU146, LYS147, ALA148, TYR177, SER179, TYR180, LEU182, LYS183, ASN184, LEU186,  
PHE278, PHE279, ILE280, PRO281, GLY282, PRO283, GLU284, GLU285, THR286, SER287,  
GLU288, VAL290, GLY293, SER294, LEU295, CYS296, ASP297, ILE300, ASP301, ILE303,  
CYS304, ILE306, LEU318

Site 2:



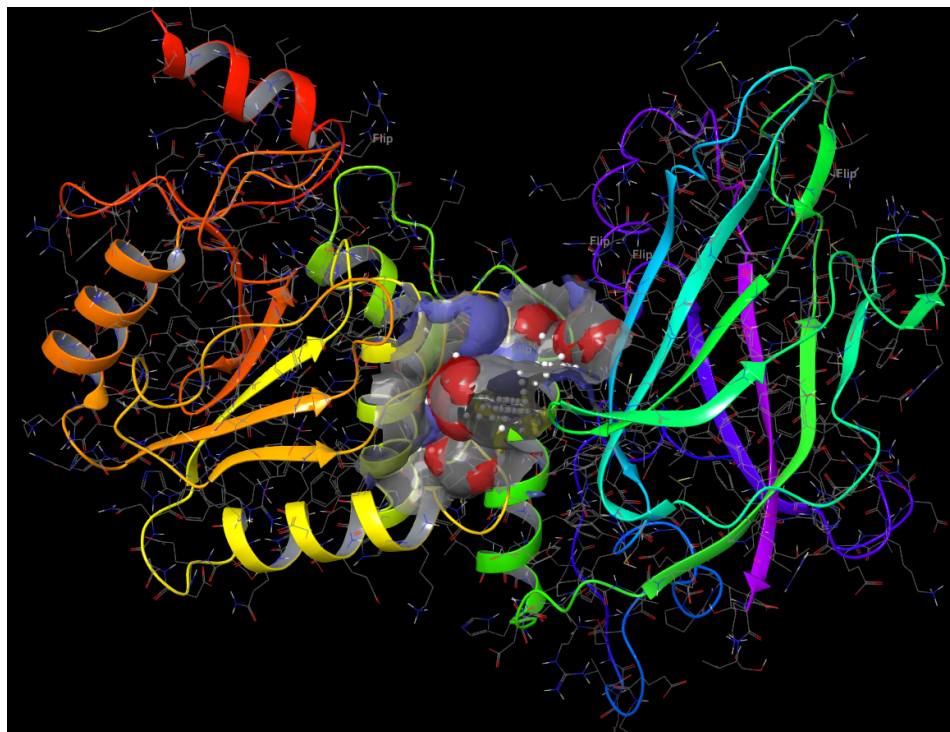
Involved residues: ARG11, ASN12, LYS13, ASP24, VAL45, TYR46, ARG47, ASP92, HSD93,  
LYS125, ALA126, LYS128, GLY129, GLY132, TYR155, THR160, GLY165, VAL166, THR167,  
ILE168, PRO169, GLN171, LYS269, PHE271, HSD272, ASP326, LYS327, ASN329, LYS330

Site 3:



Involved residues: LEU98, GLU99, LYS102, HSD141, TYR178, LEU181, LEU182, HSD185, LEU186, ASP187, TYR188

Site 4:



Involved residues: PRO89, PHE90, GLU91, ASN94, PRO95, GLN97, LEU100, CYS218, LEU220, LYS221, ASP252, LYS254, TRP274

#### Site 5:



Involved residues: PRO169, ARG172, ARG173, TYR176, GLU299, SER302, ILE303, LEU320, ASP324


#### **3. Select a cavity and create the grid where the virtual screening will take place. Explain briefly the procedure (from the manual). (1 point)**

The selection of one of the binding pockets for further analysis in this case is based on a combination of factors, including the topology of the pocket, as well as its SiteScore, which is used to quantify the potential of a binding site. The SiteScore is a measure of the overall quality of the binding site. It takes into account several factors, including the size and shape of the site, its electrostatic potential, and its accessibility to the solvent. The SiteScore ranges from 0 to 1, with higher values indicating higher quality binding sites. A SiteScore of 0.80 has been found to accurately distinguish between drug-binding and non-drug-binding sites.<sup>[2]</sup>

Based on the above, binding site 5 is chosen for further analysis, due to its good SiteScore in addition to its topology, being close to but not directly on PTEN's active site (res. 121-131)<sup>[3,4]</sup>, allowing for allosteric interactions with potential ligands.

The purpose of Glide's grid generation is to prepare a 3D grid that represents the binding site of a target protein for molecular docking simulations. The grid is used to represent the protein's electrostatic potential, which is important for accurately predicting the binding affinity of potential ligands to the protein.<sup>[5]</sup>



After selecting binding site 5, we search for and click on Glide's Receptor Grid Generation in Maestro's tasks, and a new window appears. We make sure to deselect "Pick to identify the ligand" in the Receptor tab, and then click on one of the involved residues of site 5  to place the grid, whose volume can also be expanded. We then click on run and wait until it is completed. The output grid file is generated and is ready for use.

#### **4. Perform Glide docking of the Maybridge database using the "SP" scoring function. Click on run and the calculation will start.**


The purpose of Glide's ligand docking is to predict the binding affinity of small molecule ligands to a target protein. Potential ligands are docked into the binding site of the target protein using the generated grid. During this simulation, the ligand's conformation is optimized to minimize its interaction energy with the protein. The interaction energy between the ligand and the protein is calculated and used to rank the ligands based on their predicted binding affinity.

The scoring function used in Glide is called the SP score (assigns a numerical score to each pose, with higher scores indicating a better fit between the ligand and receptor), which takes into account a variety of factors including the ligand's binding energy, its hydrogen bonding interactions with the protein, and its hydrophobic interactions.

The results of the docking simulation are analyzed to identify the most promising ligands for further study. The predicted binding affinity of the ligands can be used to guide the design of new drugs or to prioritize potential drug candidates for further experiments.<sup>[5]</sup>

After performing the Grid Generation, we search for and click on Glide's Ligand Docking in Maestro's tasks and a new window appears. In the "Ligand docking" panel, under "Receptor Grid, Filename:" we browse for the grid file we generated. Below that, at the "Input ligands from: Files" field, we browse for the file "MaybridgeHitfinder.sdf", which is the original Maybridge database (a drug library containing 24,000 compounds), that we want to virtual screen against our protein structure. We then click on run and wait until the job finishes. Finally, we save the first 1000 top SP Glide compounds (with the lowest docking score) from the screening in an sdf file, by first selecting them in our workspace and using Maestro's Extract Structures option.

#### **5. Use the ChemBioServer to filter your 1,000 top-ranked compounds for bad vdW interactions. (1 point)**

ChemBioServer 2.0 is a website that hosts a group of tools and web services developed to provide advanced filtering, clustering and networking of chemical compounds, facilitating both drug discovery and repurposing. One of those tools is the van der Waals filter  using distance and energy tests, which filters out the compounds containing bad vdW interactions (steric clashes by means of van der Waals energy and radii tolerance).<sup>[7]</sup>

We click on the Filtering → Van der Waals tab, upload our previously generated sdf file to the server, use the default vdW parameters and click on "Process Data". After successful execution, we download a new sdf file that contains the compounds that passed the vdW test, and can use it for further analysis.

**6. Filter the compounds that passed the bad vdW filter for toxic moieties using the ChemBioServer. (1 point)**

Another tool that ChemBioServer provides is the toxicity filtering using specific organic toxic roots, which filters out the compounds containing any of those undesired toxic moieties.<sup>[7]</sup>

We click on the Filtering → Toxicity tab, upload our previously generated sdf file to the server and click on “Process Data”. After successful execution, we download a new sdf file that contains the compounds that passed the toxicity test, and can use it for further analysis.

**7. Use Qikprop (Maestro) to calculate solubility (QPlogS), cell permeability (QPCaco), and number of metabolites of compounds that have passed through all the filters. Filter your compounds for QPlogS > -6.5, QPCaco > 22 nm<sup>2</sup>/s, #metabolites <7. (1 point)**

QikProp is another software tool that comes with Maestro, which is used to predict various properties of small molecule compounds, such as pharmacokinetic and pharmacodynamic parameters, as well as toxicity and ADME (Absorption, Distribution, Metabolism, and Excretion) properties.

The basic principle behind QikProp is the use of molecular descriptors, which are mathematical representations of molecular structure. These descriptors are used as inputs to a machine learning model, which is trained on a large dataset of compounds and their corresponding properties. QikProp uses the molecular descriptors and machine learning model to predict the properties of new compounds, without the need for time-consuming and expensive experimental measurements. This makes it a useful tool for rapidly screening large libraries of compounds to identify promising drug candidates.<sup>[6]</sup>

We load our latest sdf file, which contains all the filtered structures, into Maestro. In Maestro's tasks, we search for QikProp and click on it, which opens a new window. We make sure to select all of the filtered structures from our project table and then click on run and wait for the job to complete. We then open our project table and sort out compounds for QPlogS > -6.5, QPCaco > 22 nm<sup>2</sup>/s and #metabolites <7 consecutively, while deleting the structures that don't meet these criteria in between each step. Finally, we select the remaining structures and export them into a new sdf file for further analysis.

**8. Bonus (+1):**

**Using the rcdk library, calculate logP. Further filter your compounds by selecting logP < 5 and plot for the remaining compounds logP against TPSA. Do you see any correlation? (you can also investigate other correlations on your own if you want)**

### 9. Cluster the remaining compounds using the clustering feature of ChemBiorver. (1 point)

Similarity search is a method of comparing the chemical elements, molecules, or compounds with respect to their structural or functional qualities, to predict their properties and to conduct drug design studies. The Similar Property Principle states that similar compounds have similar properties, so molecules that are located closely together in chemical reference space are often considered to be functionally related.

In the context of filtering a drug library for potential ligands on a specific protein, hierarchical clustering is a technique used to group similar molecules together. The dissimilarity between the molecules is measured using a distance metric, such as Euclidean distance or Tanimoto index, based on molecular descriptors, fingerprints, or structural keys. The purpose of using hierarchical clustering is to select different compounds from a given population and to evenly populate a given chemical space with candidate molecules. The Tanimoto coefficient is a commonly used similarity metric that compares the presence or absence of structural fragments in two molecules.

Hierarchical clustering is performed on a set of molecular descriptors to form a tree-like structure that represents the relationships between the molecules based on their dissimilarity. The most populated clusters are then chosen for further investigation as they represent the most diverse set of compounds that have similar properties.<sup>[7,8]</sup>

Another tool that ChemBioServer provides is one that performs the aforementioned hierarchical clustering of molecules. We click on the Clustering → Hierarchical tab, upload our previously generated sdf file to the server, select Soergel (Tanimoto Coefficient) as the distance method, Ward Linkage as the clustering method, 0.99 as the clustering threshold and click on "Process Data". After successful execution, we got 188 clusters of similar compounds. We pick the 5 most populated clusters and download their sdf files, so that we can use them for further analysis.

### 10. Calculate the exemplars from your clusters. (1 point)

Affinity Propagation Clustering is a method for clustering data into multiple groups or clusters, with the aim of finding dense, populated clusters of similar data points. The algorithm works by defining a similarity or "affinity" measure between data points, and using this to propagate information about the preferred cluster assignments of each data point, until convergence is achieved. The result of the algorithm is a set of exemplars, one for each cluster, which are chosen to be representative of the cluster as a whole.

Exemplars are chosen from among the data points to represent each cluster. These exemplars are typically chosen to be those data points with the highest sum of affinities to all the other points in the cluster. The exemplars serve as a compact representation of each cluster, and can be used for further analysis, such as visualizing the clusters, or for modeling the underlying relationships between the data points.<sup>[7,8]</sup>



Another tool that ChemBioServer provides is the aforementioned affinity propagation clustering, providing exemplars for each cluster. We click on the Clustering → Affinity Propagation tab, upload our previously generated sdf files to the server one by one, select Soergel (Tanimoto Coefficient) as the distance method and click on “Process Data”. After successful execution, we got 2-4 exemplar compounds from each cluster. We download each sdf file, so that we can use them for further analysis.

**11. The post-processing of your virtual screening results resulted in a few compounds that may be purchased for biological assaying. Visualize your results and select the five most promising compounds from your screening and postprocessing. The criteria that you need to use to select these compounds are:**

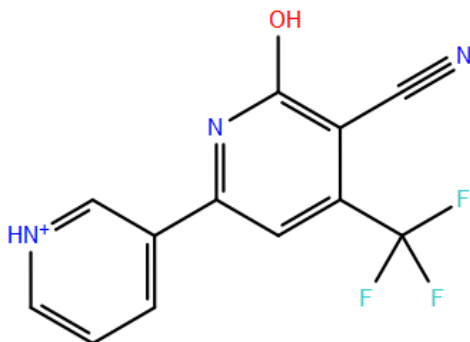
**-no more than 10 rotatable bonds**

**-none or 1 chiral centers**

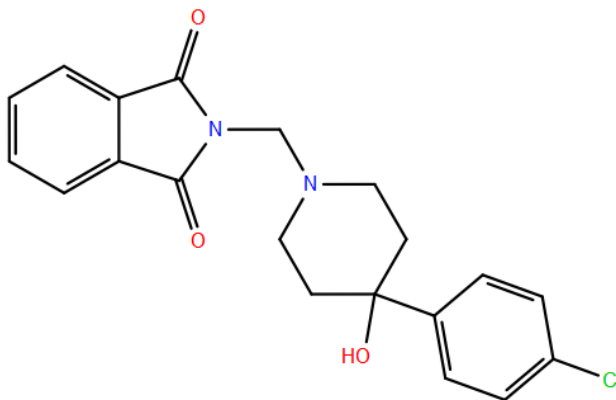
**(1 point)**

We load the 5 sdf files containing our exemplar compounds into Maestro. To select the 5 most promising compounds from our screening and postprocessing, we first open the project table and filter out any structures with more than 10 rotatable bonds. From the remaining compounds, we visualize them to identify the ones with none or 1 chiral centers (atoms with 4 different atoms/groups attached to them). We ended up with the following 5 compounds:

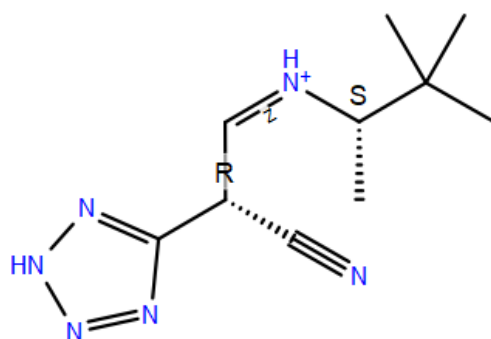
RH 02165



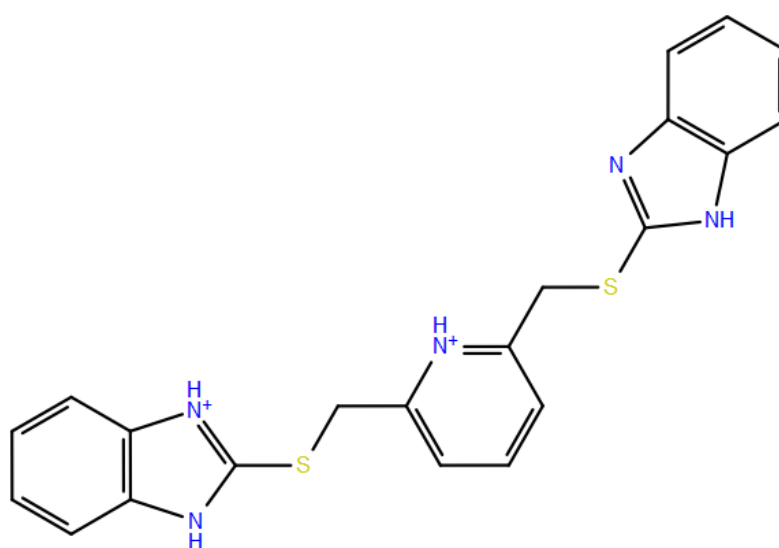
HTS 00735



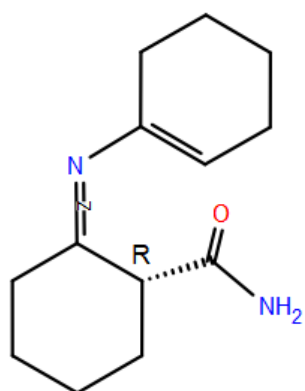
CD 11595



BTB 14890



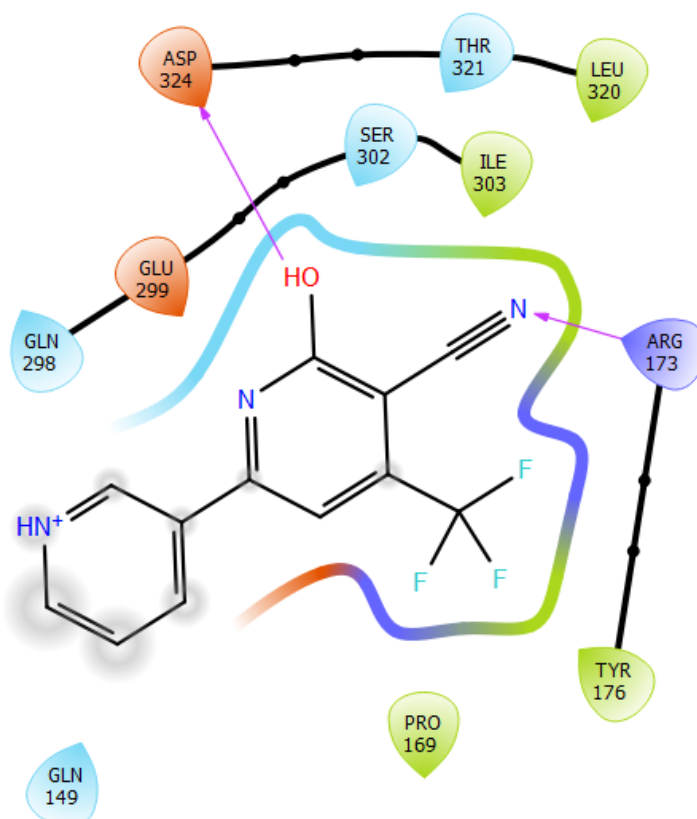
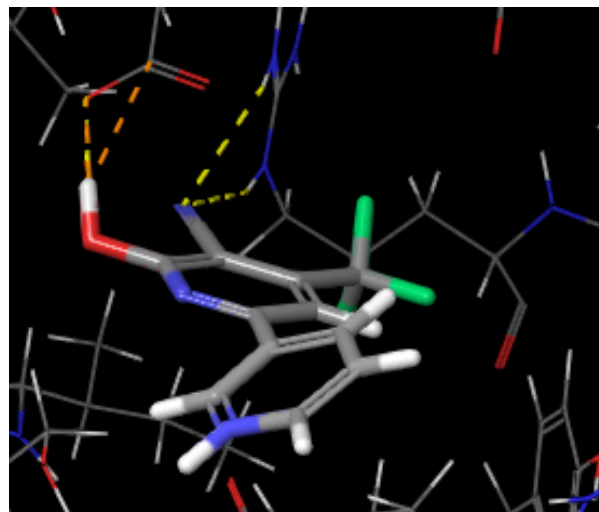
JFD 03877



12. For five of these compounds describe the intermolecular interactions between the compound and your protein. (1 point)

RH 02165

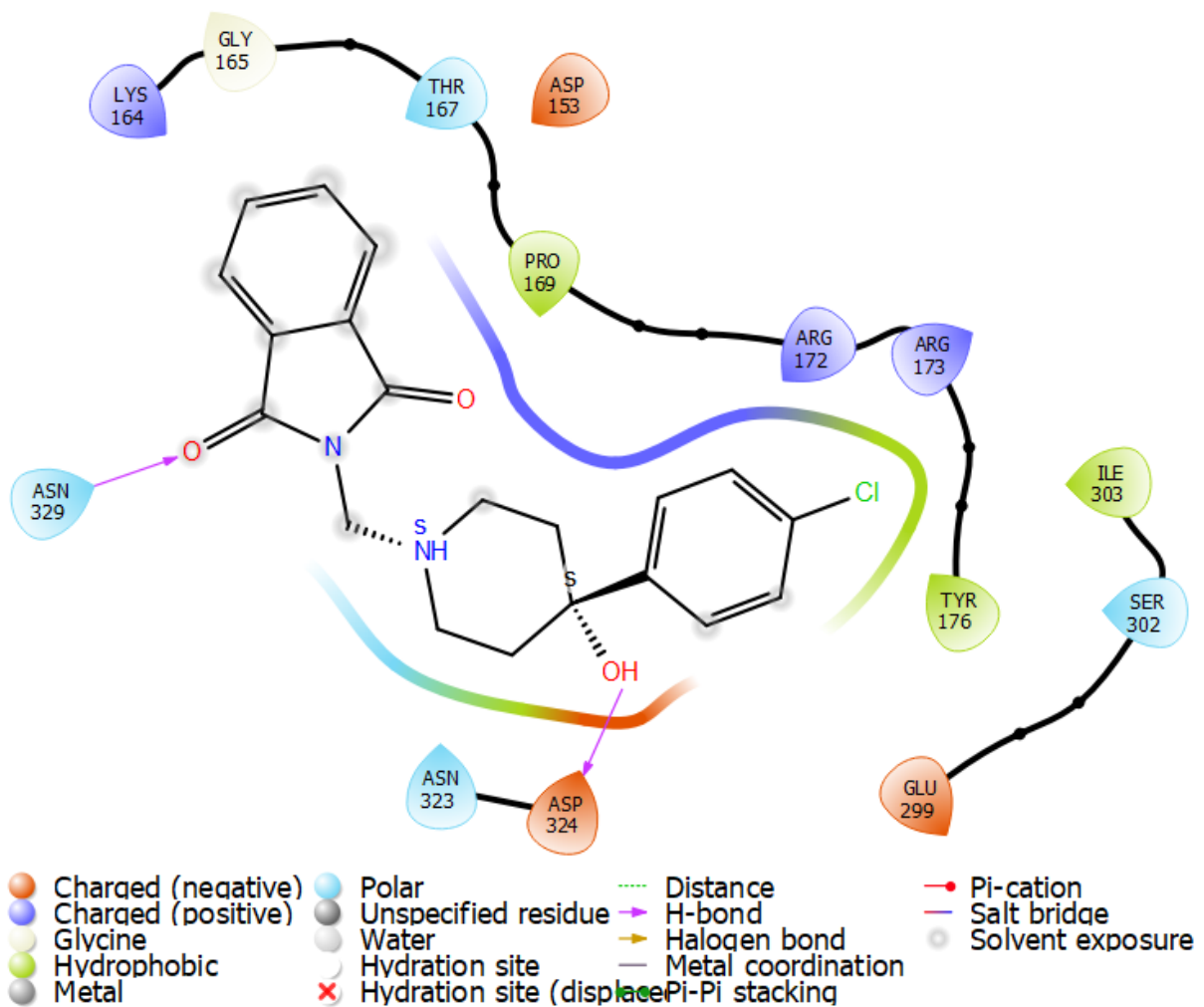
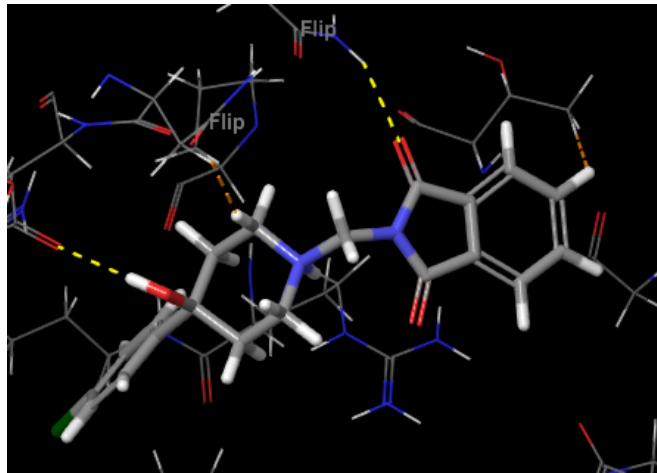
Intermolecular interactions: 3 hydrogen bonds and 2 bad contacts



- |                      |                              |                      |                    |
|----------------------|------------------------------|----------------------|--------------------|
| ● Charged (negative) | ● Polar                      | ..... Distance       | → Pi-cation        |
| ● Charged (positive) | ● Unspecified residue        | → H-bond             | → Salt bridge      |
| ● Glycine            | ● Water                      | → Halogen bond       | ○ Solvent exposure |
| ● Hydrophobic        | ● Hydration site             | → Metal coordination |                    |
| ● Metal              | ✗ Hydration site (displaced) | → Pi-Pi stacking     |                    |

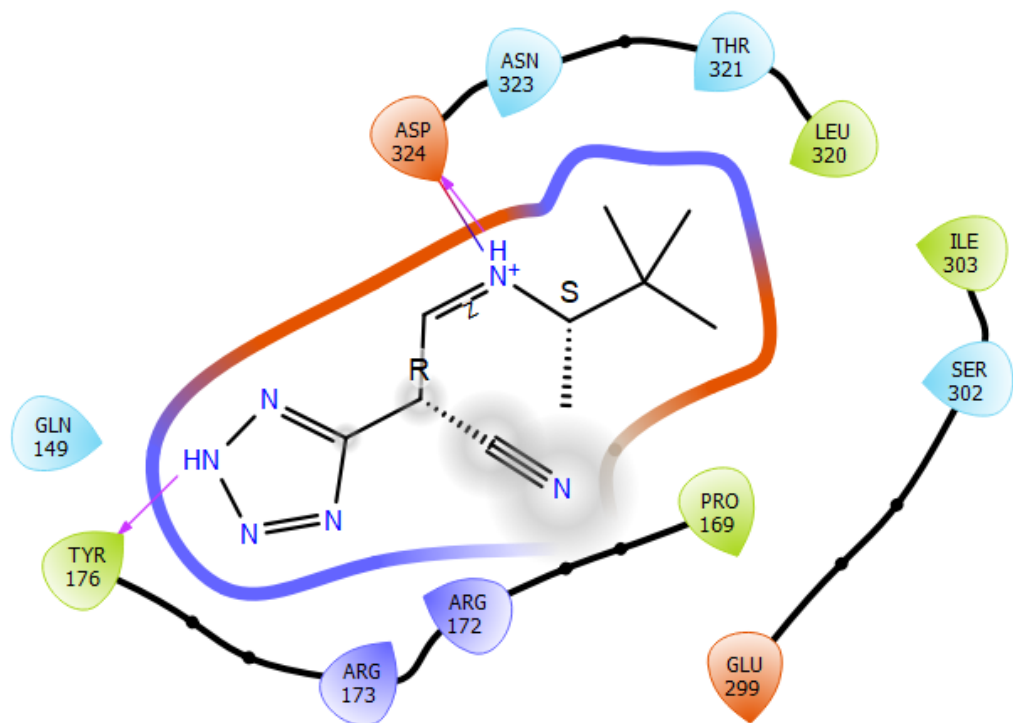
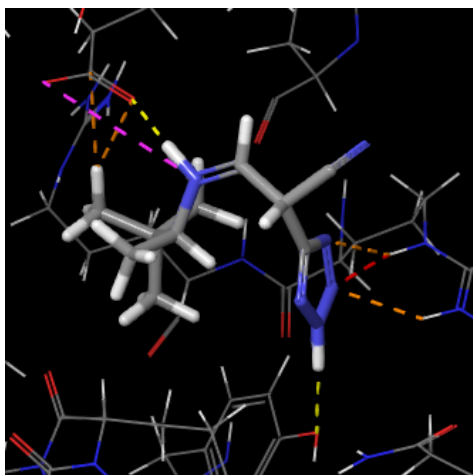
HTS 00735

Intermolecular interactions: 2 hydrogen bonds and 2 bad contacts



# CD 11595

Intermolecular interactions: 2 hydrogen bonds, 1 salt bridge, 4 bad contacts and 1 ugly contact

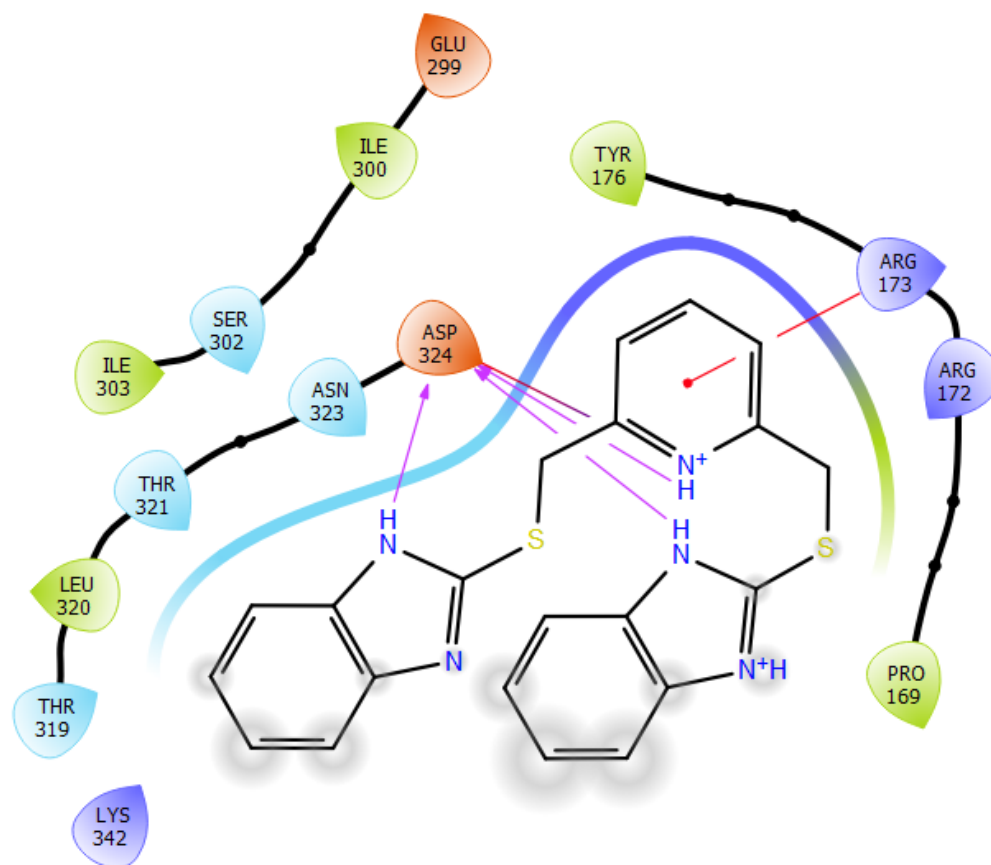
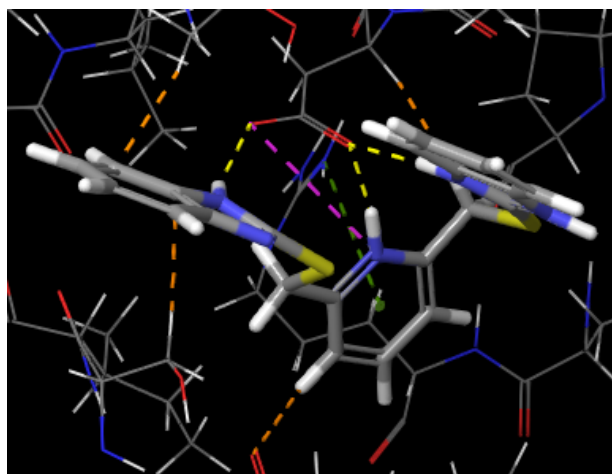


- |                      |                              |                      |                    |
|----------------------|------------------------------|----------------------|--------------------|
| ● Charged (negative) | ● Polar                      | --- Distance         | — Pi-cation        |
| ● Charged (positive) | ● Unspecified residue        | → H-bond             | — Salt bridge      |
| ● Glycine            | ● Water                      | → Halogen bond       | ○ Solvent exposure |
| ● Hydrophobic        | ○ Hydration site             | — Metal coordination |                    |
| ● Metal              | ✗ Hydration site (displaced) | — Pi-Pi stacking     |                    |



# BTB 14890

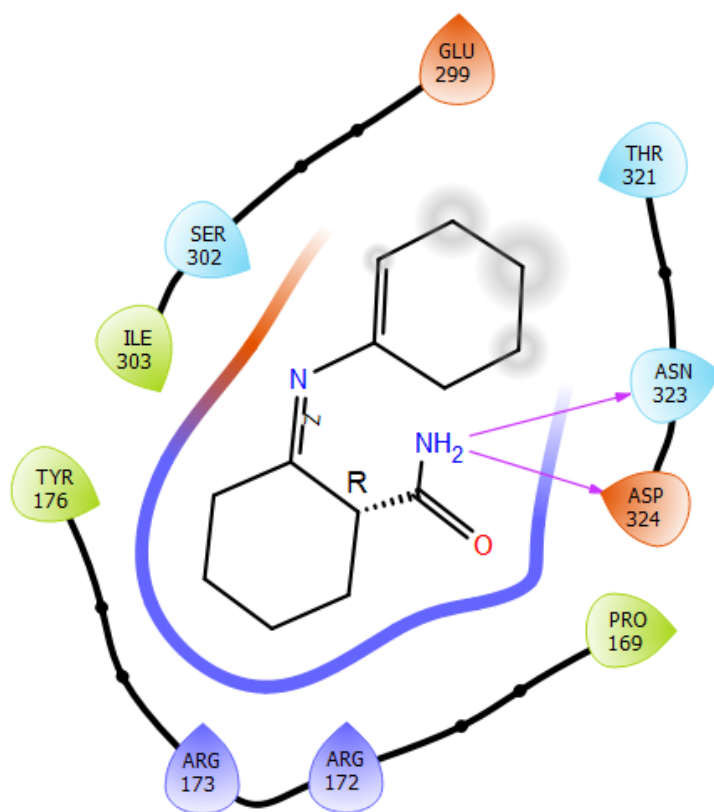
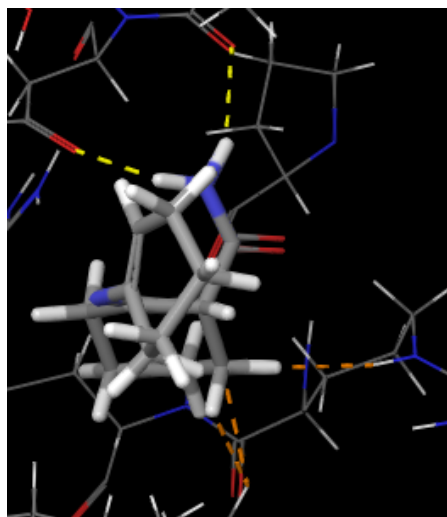
Intermolecular interactions: 3 hydrogen bonds, 1 salt bridge, 1 cation- $\pi$  interaction and 4 bad contacts



- |  |   |  |  |
|--|---|--|--|
| <span style="color: orange;">●</span> Charged (negative) | <span style="color: lightblue;">●</span> Polar                | <span style="color: green;">---</span> Distance            | <span style="color: red;">●</span> Pi-cation         |
| <span style="color: blue;">●</span> Charged (positive)   | <span style="color: grey;">●</span> Unspecified residue       | <span style="color: purple;">---</span> H-bond             | <span style="color: red;">---</span> Salt bridge     |
| <span style="color: yellow;">●</span> Glycine            | <span style="color: lightgrey;">●</span> Water                | <span style="color: brown;">---</span> Halogen bond        | <span style="color: grey;">○</span> Solvent exposure |
| <span style="color: green;">●</span> Hydrophobic         | <span style="color: white;">○</span> Hydration site           | <span style="color: purple;">---</span> Metal coordination |  |
| <span style="color: grey;">●</span> Metal                | <span style="color: red;">✗</span> Hydration site (displaced) | <span style="color: green;">---</span> Pi-Pi stacking      |  |

JFD 03877

Intermolecular interactions: 2 hydrogen bonds and 3 bad contacts



- |                    |                            |                    |                  |
|--------------------|----------------------------|--------------------|------------------|
| Charged (negative) | Polar                      | Distance           | Pi-cation        |
| Charged (positive) | Unspecified residue        | H-bond             | Salt bridge      |
| Glycine            | Water                      | Halogen bond       | Solvent exposure |
| Hydrophobic        | Hydration site             | Metal coordination |                  |
| Metal              | Hydration site (displaced) | Pi-Pi stacking     |                  |

The optimal quality of an exemplar in the context of intermolecular interactions with a protein would depend on the specific goals of the drug design study and the properties of the protein. In general, having more intermolecular interactions with the protein can suggest a stronger binding affinity, which is often desirable in drug design studies. Solvent exposure, on the other hand, may indicate that the compound is less stable or less likely to remain in its active conformation, which may negatively impact its potential as a drug. However, in some cases, solvent exposure can be desirable, for example, if the goal is to design drugs that can easily diffuse into tissues. Therefore, whether solvent exposure is considered good or bad would depend on the specific goals and context of the drug design study.

To evaluate these intermolecular interactions with our protein and judge their quality as potential ligands, we can use molecular docking techniques, which predict the binding mode of the molecule with the protein based on the 3D structures of both. The output of the docking simulation is a score that reflects the affinity of the molecule for the protein, with higher scores indicating stronger binding. We can use these scores to rank the exemplar compounds and choose the ones with the highest affinity for further evaluation and testing. Additionally, we can use molecular dynamics simulations to study the dynamics of the molecule-protein complex and assess the stability of the binding. These simulations can also provide information on other important properties, such as ligand-protein binding free energy, the nature of the interactions between the molecule and the protein, and changes in protein conformation upon ligand binding.



### References:

1. Schrödinger, Maestro 10.2 User Manual, 2015, PDF.
2. Schrödinger, SiteMap 3.5 User Manual, 2015, PDF.
3. Lee, Jie-Oh, et al. "Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association." *Cell* 99.3 (1999): 323-334.
4. Kotzampasi, Danai Maria, et al. "The orchestrated signaling by PI3K $\alpha$  and PTEN at the membrane interface." *Computational and Structural Biotechnology Journal* (2022).
5. Schrödinger, Glide 6.7 User Manual, 2015, PDF.
6. Schrödinger, QikProp 4.4 User Manual, 2015, PDF.
7. Biomedical Research Foundation, Academy of Athens, The Cyprus Institute of Neurology and Genetics (2011, December 30). ChemBioServer 2.0. <https://chembioserver.vi-seem.eu/index.php>.
8. Cournia, Z. (2023). Principles of Computer-Aided Drug Design [PowerPoint presentation]. Molecular Modeling of Biomolecules course, DSIT 2022-2023.