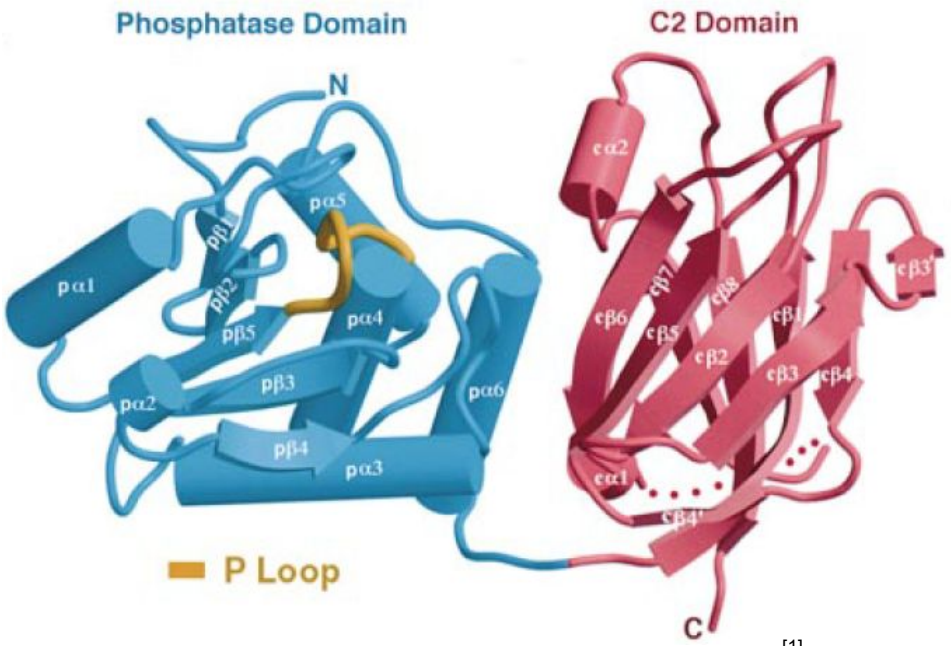
Molecular Dynamics Simulation and Computer-Aided Drug Design for the PTEN Protein

Konstantinos Giatras

Course: Molecular Modeling of Biomolecules
DSIT 2022-2023

The PTEN Protein

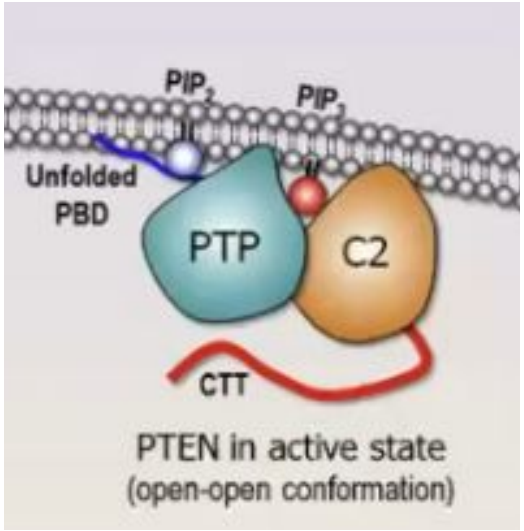
- PTEN is a protein critical for maintaining cell homeostasis, regulating metabolism and acts as a tumor suppressor in humans.
- The PTEN gene is located on chromosome 10 and is mutated in many human cancers and cancer predisposition syndromes.
- PTEN has both protein phosphatase and phosphoinositide phosphatase activity *in vitro* and is able to remove phosphate groups from intracellular phosphoinositide signaling molecules.
- PTEN has potential as a therapeutic target for tissue regeneration, nerve injuries, and Alzheimer's disease.



[1]

The PTEN Protein

- The structure of the PTEN protein includes, among others, a phosphatase domain and C2 domain.
- The active site of PTEN between the two domains is important for accommodating the phosphoinositide (PIP_3) substrate → PIP_2.
- PIP_2 is needed for the PTEN protein to be able to bind to the membrane and be functional.
- PI3Kα: $PIP_2 \rightarrow PIP_3$

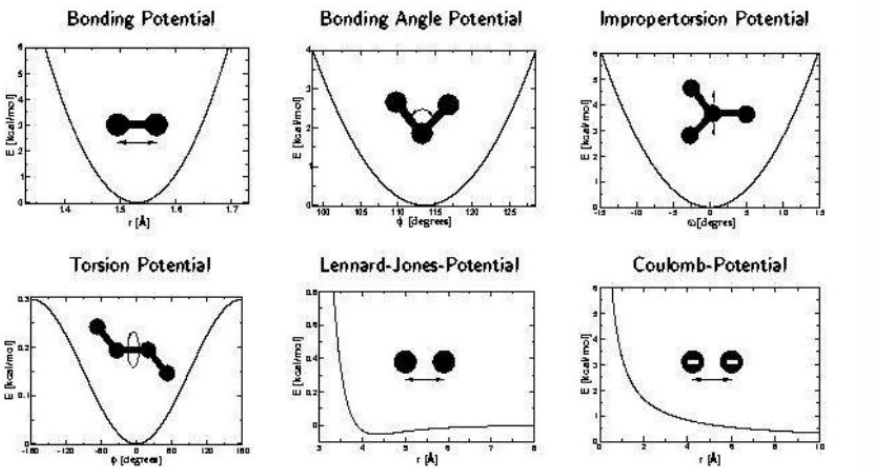


[2]

Molecular Dynamics (MD) Simulations

- Computational methods that use numerical integration of Newton's equations of motion to simulate the motion of atoms and molecules over time.
- Predictions of the structure, dynamics, and thermodynamics of molecular systems.
- CHARMM force field: set of parameters that describe the interactions between atoms and molecules, including bonded and nonbonded interactions such as covalent bonds, angles, dihedrals, van der Waals and electrostatic interactions → Potential Energy.
- Intermolecular interactions (between molecules): can be either covalent or noncovalent, including electrostatic, van der Waals, hydrogen bonds, π-π and cation-π interactions → essential for determining the behavior and properties of molecules.

CHARMM Potential Energy Function (Force Field)



$$E = \frac{1}{2} m\mathbf{v}^2 + V(\mathbf{r}) \qquad \mathbf{F}_i = -\nabla V(\mathbf{r}) \qquad V(\mathbf{r}) = E_{bonded} + E_{non-bonded}$$

$$E_{bonded} = \sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \delta]) + \sum_{impropers} k_\omega (\omega - \omega_0)^2$$

$$E_{non-bonded} = \sum_{i,j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i,j} \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}}$$

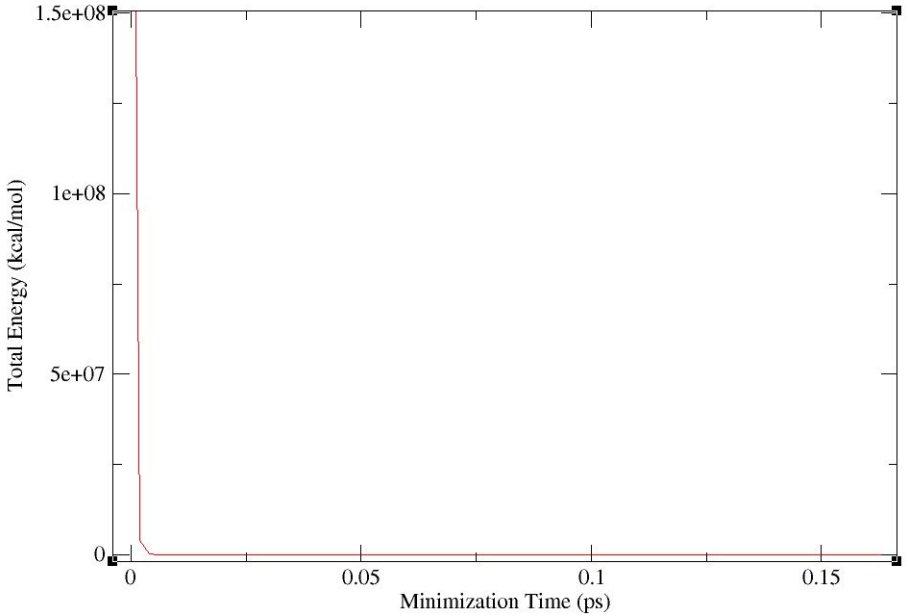
[7]

Resources:

- Protein Data Bank (PDB): repository of experimentally determined three-dimensional structures of biological macromolecules, including proteins, nucleic acids, and complex assemblies.
- VMD: molecular visualization program that is used to display and analyze molecular dynamics simulations and other biomolecular structures.
- NAMD: parallel molecular dynamics simulation software designed for high-performance simulation of large biomolecular systems.
- XMGrace: plotting and graphing tool used for visualizing scientific data, particularly in the field of physics and chemistry.

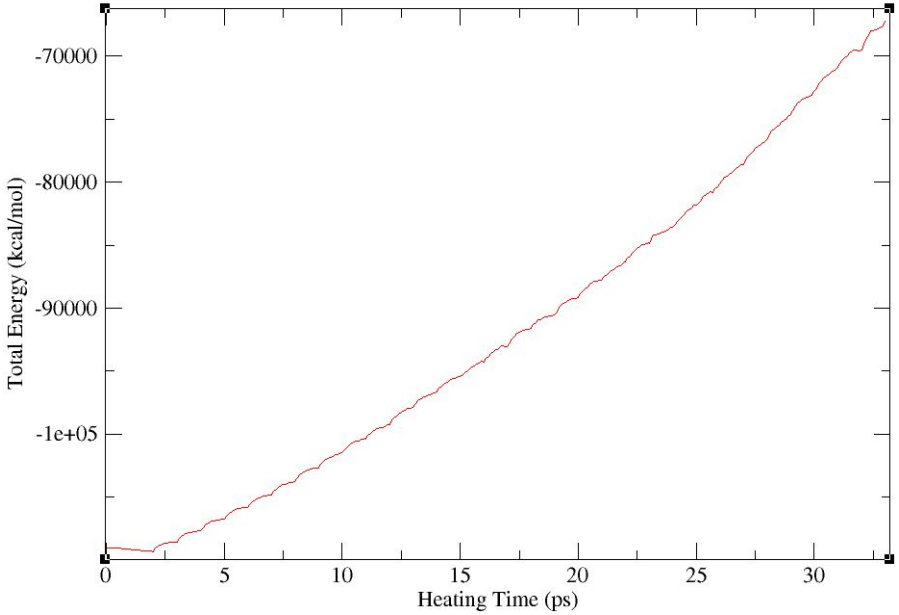
Protein Preparation and Minimization

- Obtain topology and parameter files, protein structure and coordinate files, and remove crystal waters if necessary.
- Use VMD to generate psf file, add missing hydrogen atoms, solvate protein, neutralize the system, and add NaCl to represent biological environment.
- Minimization (remove any potential clashes and minimize the energy of the initial structure): configure input file and run namd2 with periodic boundary conditions and 5000 minimization steps, obtain log file and grep 'ENERGY:' into dat file, plot Total Energy vs Minimization Time using XMGrace.
- Result: plot shows that the minimization has converged.



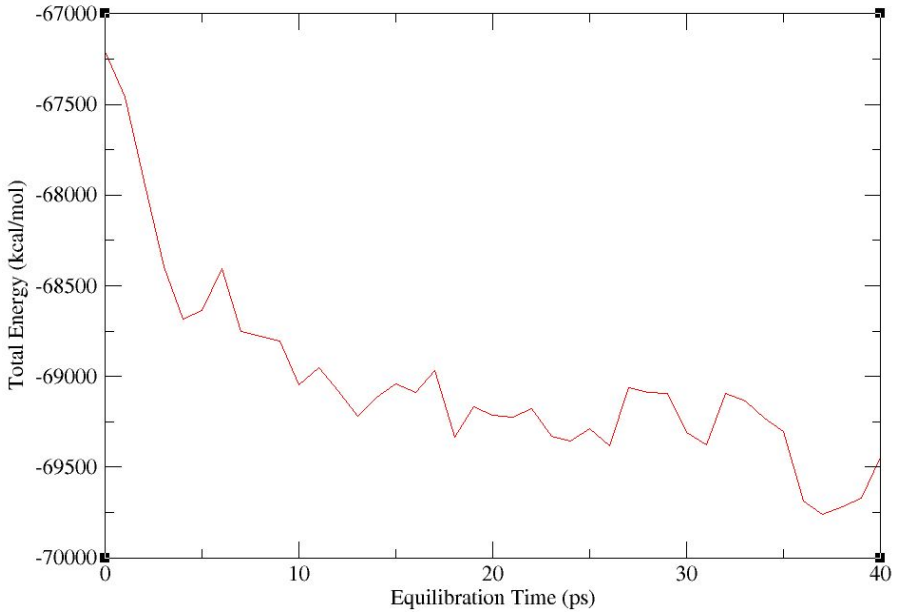
Heating

- The heating phase involves assigning initial velocities to each atom in the system at low temperature and integrating using Newton's equations of motion.
- The simulation is initiated by assigning velocities at slightly higher temperatures in a periodic manner until the target temperature is reached.
- This process is repeated with progressively higher initial velocities.
- The heating is run for 500 simulation steps every 10 K, starting from 0 K and ending at 310 K, and generates a log and then a dat file.
- XMGrace is used to plot the data (Total Energy vs Heating Time).



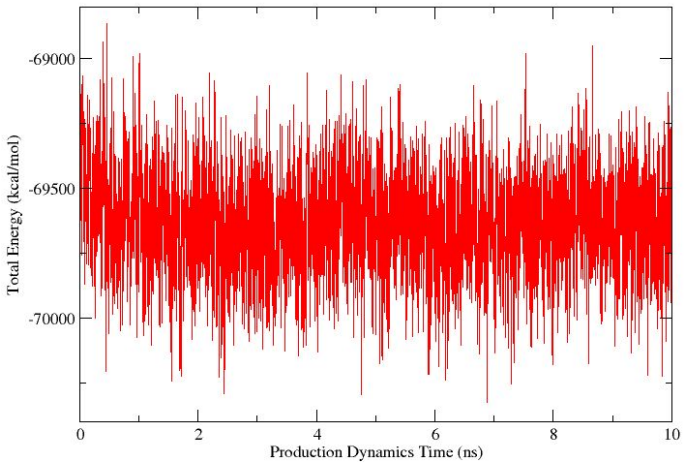
Equilibration

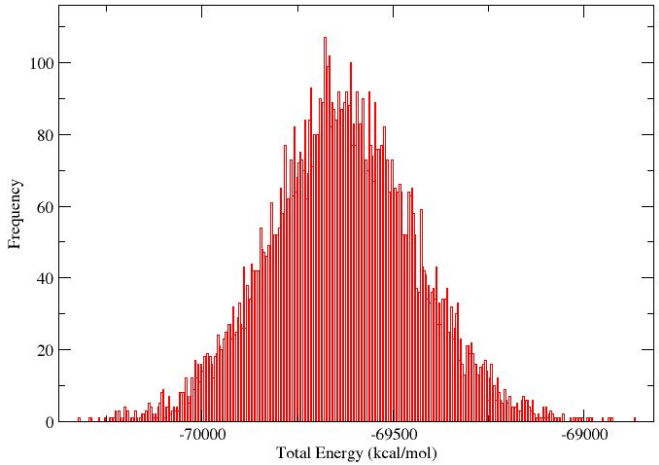
- Equilibration involves solving Newton's Second Law for each atom to determine its trajectory using molecular dynamics
- Stability of properties such as structure, pressure, temperature, and energy with respect to time are monitored.
- If temperature deviates significantly from the desired value, velocities are scaled to bring the temperature back to the target value.
- After running the equilibration and generating the dat file, XMGrace is used to plot the data (Total Energy vs Equilibration Time).
- The system appears to have reached thermodynamic equilibrium, as the plot has reached a plateau, and therefore there is no need to restart the equilibration.

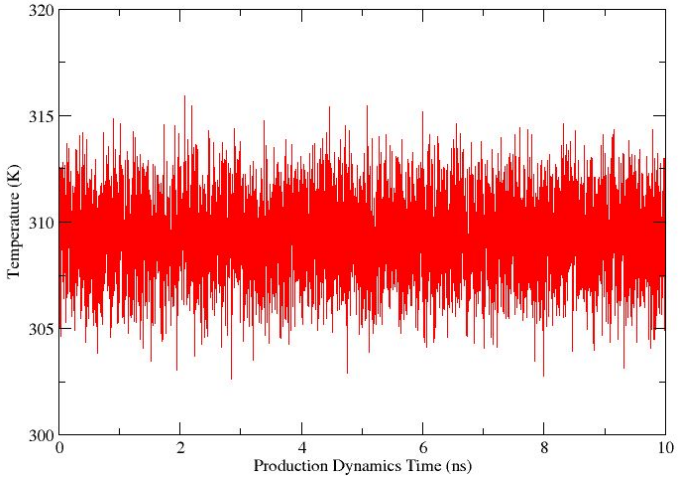


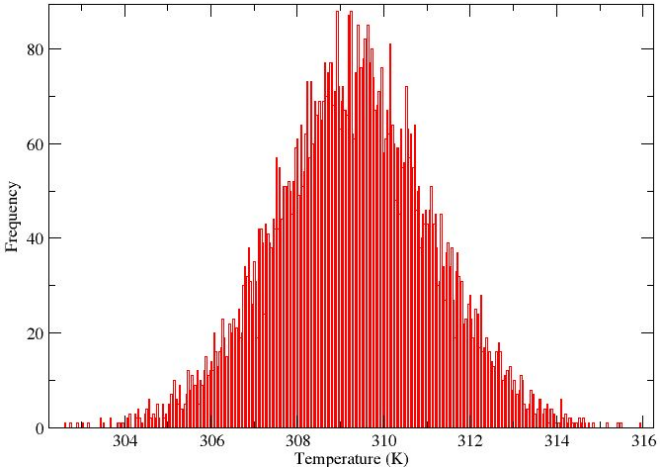
Productions Dynamics

- Production dynamics is the final step of a simulation where a system is simulated for a specific time period.
- The purpose is to collect data, record trajectories, and calculate various properties such as mean energy and root mean square deviation (RMSD).
- The data collected is plotted using XMGrace. The x-axis shows the production dynamics time steps and the y-axis shows the total energy and temperature.
- Fluctuations observed are within statistical variance occurring at equilibrium, which means the system remains stable throughout the production dynamics run.
- The histograms generated have the functional form of Gaussian distributions, which occur when the system is in equilibrium.


Total Energy (kcal/mol)
-69000
-69500
-70000
0
2
4
6
8
10
Production Dynamics Time (ns)


Frequency
100
80
60
40
20
0
-70000
-69500
-69000
Total Energy (kcal/mol)


Temperature (K)
320
315
310
305
300
0
2
4
6
8
10
Production Dynamics Time (ns)

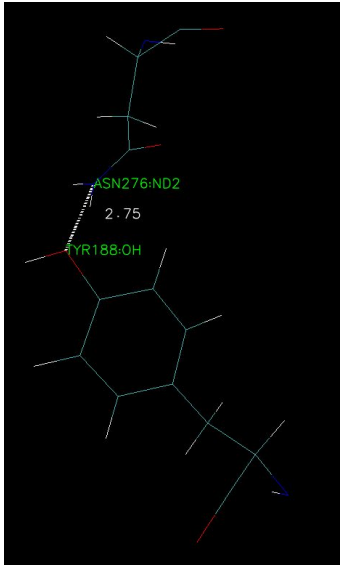

Frequency
80
60
40
20
0
304
306
308
310
312
314
316
Temperature (K)

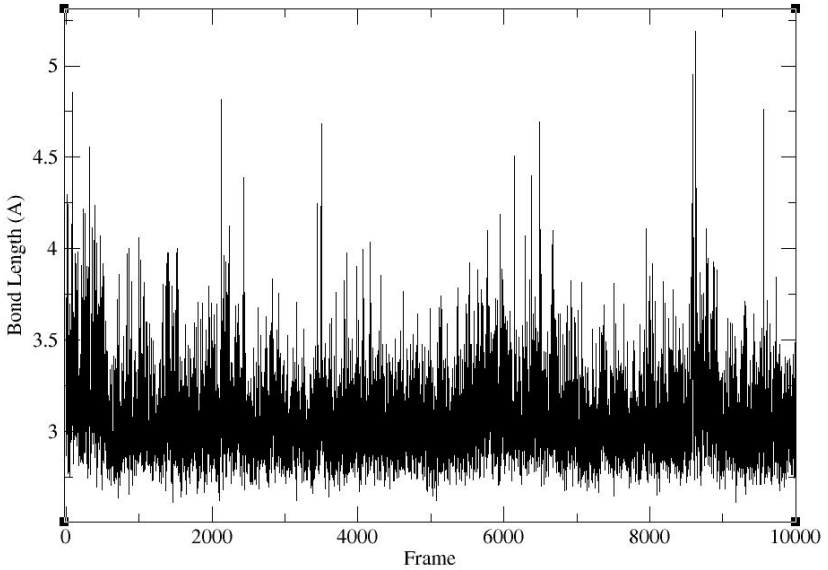
Trajectory Analysis

- Trajectory analysis involves examining movements and changes in protein conformation during a simulation.
- Properties calculated can include RMSD (measure of the differences between the conformations of a protein over time, which is calculated by comparing the positions of the atoms in the protein structures in a trajectory to a reference structure), residue mobility, and conformational changes.
- The resulting data can provide insights into protein function and drug design.

Trajectory Analysis

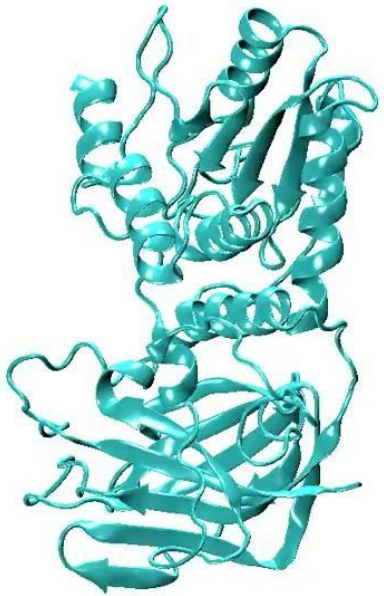
- The task at hand is to identify the parts between N276 and Y188 that form a hydrogen bond and plot its time series throughout the trajectory.
- The nitrogen from asparagine's side chain forms a hydrogen bond with the oxygen from tyrosine's side chain.
- The hydrogen bond is characterized by both bond distance and angle.
- The average distance (3.068 Å) and standard deviation (0.231 Å) of the bond are calculated using Excel.
- The hydrogen bond distance criterion (on average 2.7-3.3 Å) is mostly satisfied throughout the trajectory.





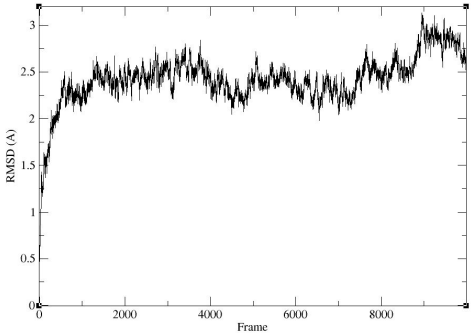
Trajectory Analysis

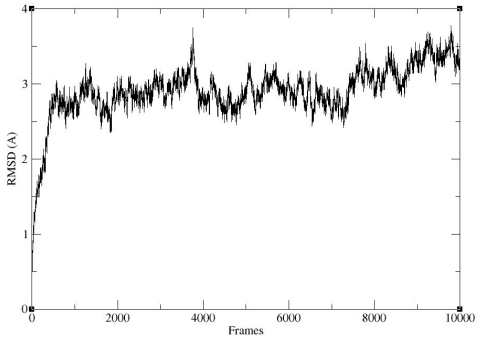
- VMD's Movie Maker is used to produce a movie of the production dynamics simulation.
- Movie Maker can be found in the Extensions → Visualization menu.
- Colors and display are configured to achieve the desired visual result.
- The Movie Maker is set up with the following settings:
  - Movie Settings: Trajectory
  - Format: ppmtompeg
  - Trajectory step size: 7
- The video is rendered and later converted to mp4.

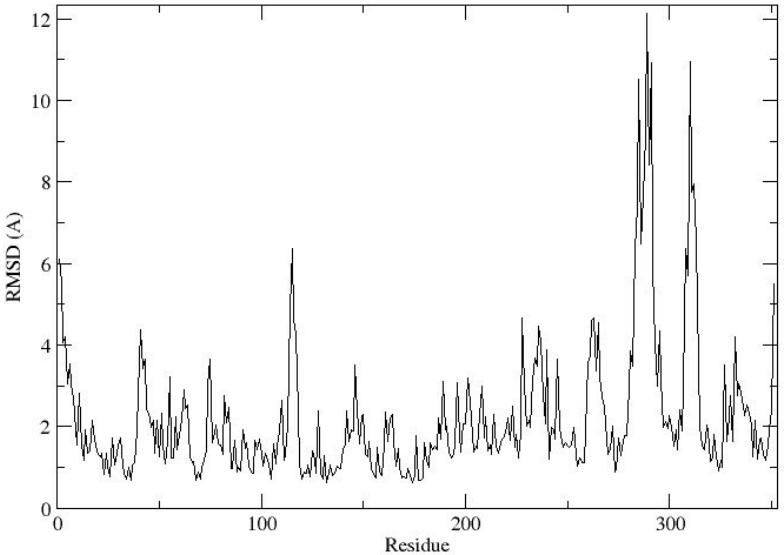


Trajectory Analysis

- The RMSD of each of the 351 residues of PTEN is calculated using a script within VMD.
- The average RMSD value and standard deviation of the PTEN Residue RMSD are calculated and found to be 2.194 Å and 1.688 Å, respectively.
- Substantial mobility is observed for most of PTEN's residues, especially at the C-terminal domain.
- The RMSD time series of PTEN and the C-terminus of PTEN are plotted using the RMSD Trajectory Tool.
- The PTEN Residue RMSD plot shows the most flexible parts of the protein, and the C-terminal domain has a higher RMSD value compared to the entire protein.
- The average RMSD value and standard deviation for the PTEN RMSD Time Series and PTEN C-Terminal Domain RMSD Time Series are calculated and found to be 2.421 Å and 0.262 Å, and 2.907 Å and 0.352 Å, respectively.
- PTEN is quite mobile, especially when it comes to its C-terminus.

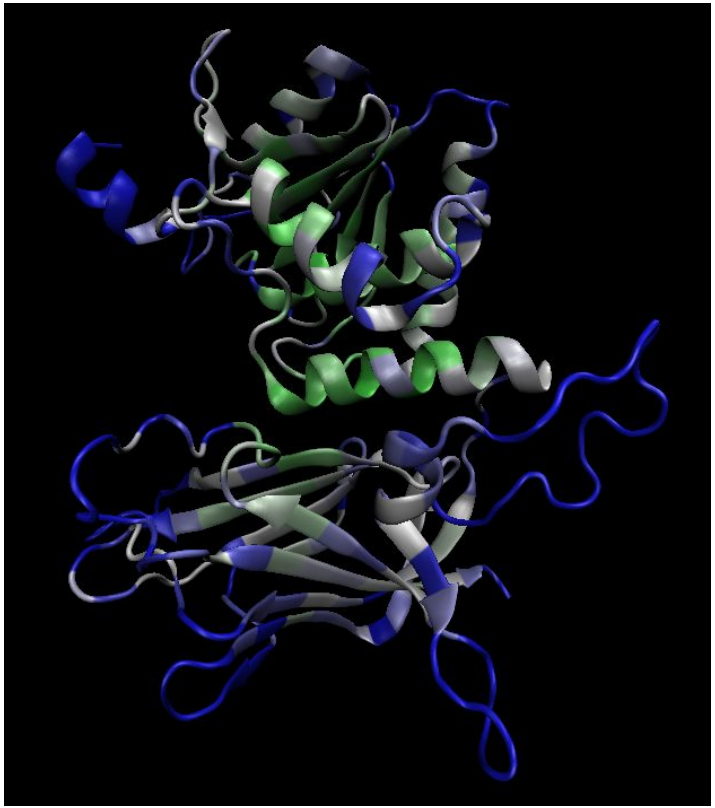






Trajectory Analysis

- The User field of all atoms in selected residues is set to the computed RMSD value
- The protein is colored in VMD according to the RMSD value, using NewCartoon as the DrawingMethod
- Residues colored in blue are more mobile while green ones move less
- Residues on the outside and loops of the protein tend to be more flexible
- Loops can act as hinges, allowing the protein to move or change shape
- This flexibility is important for the function of the protein



Trajectory Analysis

- Clustering is a method of identifying similar groups of structures in a protein trajectory.
- Hierarchical clustering creates a hierarchy of clusters, with each cluster being split into smaller clusters until each individual structure is in its own cluster.
- RMSD is used as the similarity measure to measure the deviation of atoms in each structure from the mean structure of the cluster.
- A code is written to find the 1st cluster representative from the protein trajectory using hierarchical clustering.
- The code extracts protein backbone atoms, calculates pairwise RMSDs, performs hierarchical clustering, and extracts cluster assignments.
- It then calculates the average RMSD for each cluster, finds the cluster with the lowest average RMSD, and the structure that is the centroid of that cluster, which is found to belong to frame 70 of the trajectory.
- The coordinates of the representative structure are saved as a new pdb file for further use.

Computer-Aided Drug Design

- Computer-aided drug design (CADD) uses computational methods to discover, design, and optimize new drugs.
- CADD can significantly reduce the cost and time required for experimental testing and allows for the exploration of a large number of molecules and their interactions.
- CADD can be used in various stages of drug discovery and can design drugs with higher specificity and potency, fewer side effects, and better pharmacokinetics.
- Maestro is a comprehensive software suite for computer-aided drug design and molecular modeling.
- Protein preparation is a crucial step in the drug discovery process, involving the preparation of a high-quality 3D structure of a target protein that accurately represents its active conformation.
- The protein preparation workflow in Maestro involves deleting waters, assigning bond orders, filling in missing side chains, replacing hydrogens, optimizing h-bond assignments, creating zero-order bonds, generating het states, and minimizing by converging heavy atoms.

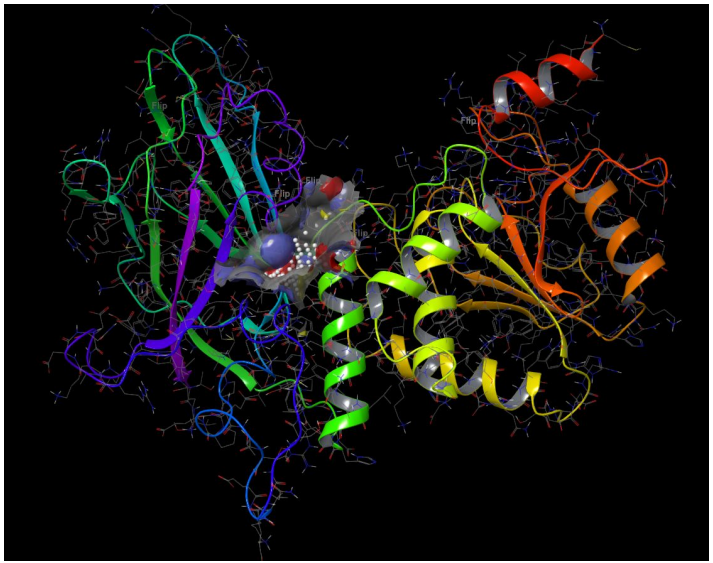
Binding Site Detection (SiteMap)

- SiteMap is a computational method in Maestro that identifies potential binding pockets on a protein structure.
- The process of SiteMap involves removing water molecules, ligands, and cofactors, dividing the protein into a grid of voxels, calculating the properties of each voxel, grouping voxels into clusters, screening clusters to identify potential binding pockets, and refining identified binding pockets using molecular dynamics simulations or other methods.
- In Maestro's tasks, SiteMap is searched for and opened in a new window after selecting the protein. The application is run with default options, and the output appears in the workspace after completion. 5 binding sites were identified.

Binding Site Detection (SiteMap)

- The selection of a binding pocket for further analysis is based on its topology and SiteScore.
- SiteScore is a measure of the overall quality of the binding site.
- SiteScore takes into account several factors, including size and shape of the site, electrostatic potential, and accessibility to solvent.
- SiteScore ranges from 0 to 1, with higher values indicating higher quality binding sites. Anything >0.8 is considered good.
- Binding site 5 is chosen for further analysis due to its good SiteScore and topology, allowing for allosteric interactions with potential ligands.

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



Grid Generation and Ligand Docking (Glide)

- Glide is a software tool in Maestro used for molecular docking and virtual screening to predict the binding of small molecules to proteins.
- Glide's grid generation prepares a 3D grid to represent the binding site of a target protein for molecular docking simulations.
- The generated grid is used to represent the protein's electrostatic potential, which is important for predicting the binding affinity of ligands to the protein.
- To create the grid, select a cavity, click on Glide's Receptor Grid Generation in Maestro's tasks, deselect "Pick to identify the ligand," place the grid on one of the involved residues of the selected site, and click on run.

Grid Generation and Ligand Docking (Glide)

- The purpose of Glide's ligand docking is to predict the binding affinity of small molecule ligands to a target protein.
- Ligands are docked into the binding site of the target protein using the generated grid, and their conformation is optimized to minimize their interaction energy with the protein.
- The SP scoring function is used to assign a numerical score to each pose, with higher scores indicating a better fit between the ligand and receptor.
- The scoring function takes into account the ligand's binding energy, hydrogen bonding interactions with the protein, and hydrophobic interactions.
- The results of the docking simulation are analyzed to identify the most promising ligands for further study.
- To perform the docking, select Glide's Ligand Docking in Maestro's tasks, browse for the generated grid file and the Maybridge database file (drug library containing 24,000 compounds), click on run, and wait until the job finishes.
- Save the first 1000 top SP Glide compounds (with the lowest docking score) from the screening in an sdf file using Maestro's Extract Structures option.

Compound Filtering using ChemBioServer

- ChemBioServer is a website with tools for filtering, clustering, and networking of chemical compounds for drug discovery and repurposing.
- One tool is the van der Waals filter, which uses distance and energy tests to filter out compounds with bad vdW interactions.
- We can upload our previously generated sdf file to the server, use the default vdW parameters, and click on "Process Data" to filter out compounds with bad vdW interactions.
- After successful execution, we can download a new sdf file containing the compounds that passed the vdW test.
- Another tool on ChemBioServer is the toxicity filter, which filters out compounds with specific organic toxic roots.
- We can upload our previously generated sdf file to the server and click on "Process Data" to filter out compounds with undesired toxic moieties.
- After successful execution, we can download a new sdf file containing the compounds that passed the toxicity test.

Further Filtering (QikProp)

- QikProp is a software tool in Maestro that predicts properties of small molecule compounds.
- It uses molecular descriptors and a machine learning model to make predictions without experimental measurements.
- QikProp can be used to screen large libraries of compounds to identify promising drug candidates.
- QikProp is used in Maestro to calculate solubility, cell permeability, and number of metabolites of compounds.
- Compounds are filtered based on QPlogS > -6.5, QPCaco > 22 nm^2/s, and #metabolites <7.
- Filtered structures are sorted and deleted based on the above criteria.
- Remaining structures are exported into a new sdf file for further analysis.

Hierarchical Clustering and Exemplar Extraction

- Similarity search compares chemical elements, molecules, or compounds based on their properties. The Similar Property Principle states that similar compounds have similar properties.
- Hierarchical clustering groups similar molecules together based on their dissimilarity using a distance metric like Euclidean distance or, in our case, Tanimoto index.
- Tanimoto coefficient is a commonly used similarity metric that compares the presence or absence of structural fragments in two molecules.

Hierarchical Clustering and Exemplar Extraction

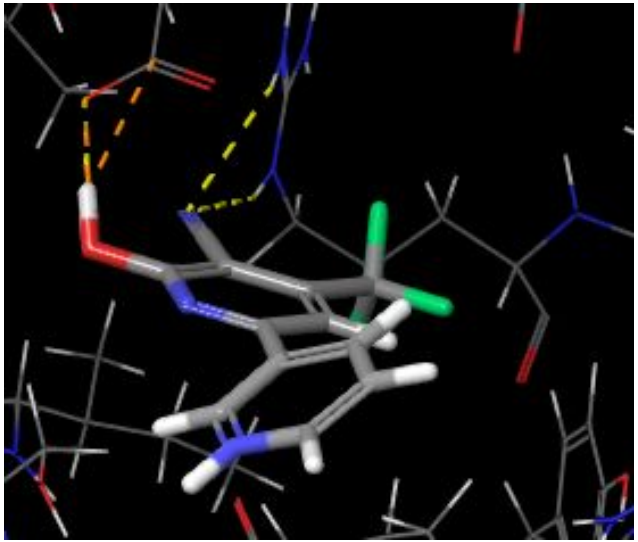
- Affinity Propagation Clustering is a method for clustering data into multiple groups or clusters to find dense, populated clusters of similar data points.
- Exemplars are chosen to represent each cluster and serve as a compact representation of the cluster as a whole.
- ChemBioServer provides tools for both hierarchical clustering and affinity propagation clustering to generate exemplars for each cluster.
- Selected Soergel as distance method, Ward Linkage as clustering method, and 0.99 as clustering threshold. Generated 188 clusters, downloaded sdf files for top 5 populated clusters and got 2-4 exemplars from each cluster.

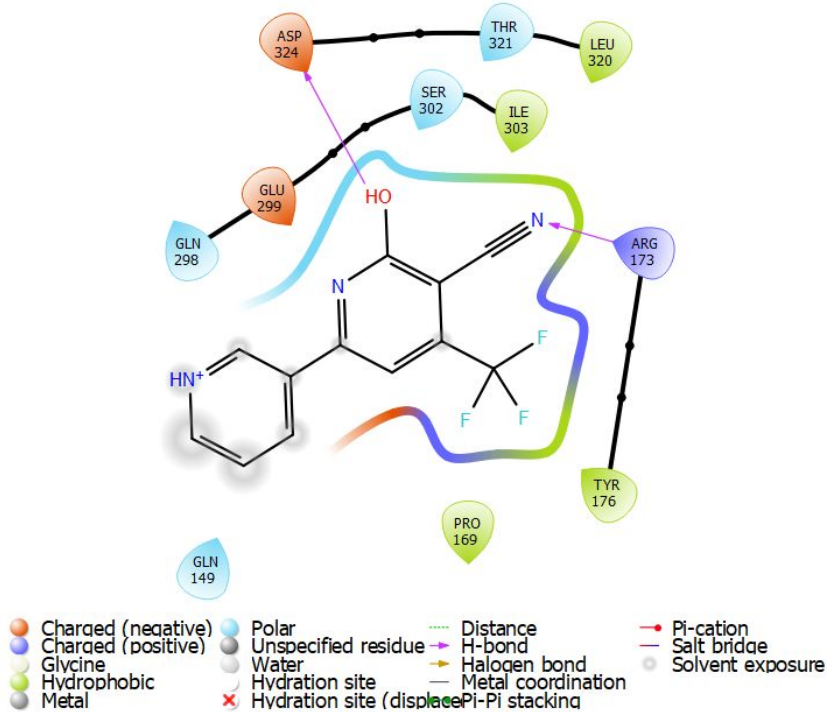
Final Compound Selection

- Five most promising compounds were selected based on the following criteria:
  - No more than 10 rotatable bonds
  - None or 1 chiral centers
- 5 sdf files containing the exemplar compounds were loaded into Maestro.
- The project table was opened and structures with more than 10 rotatable bonds were filtered out.
- Compounds were visualized to identify those with none or 1 chiral centers.
- The following 5 compounds were selected:

RH 02165

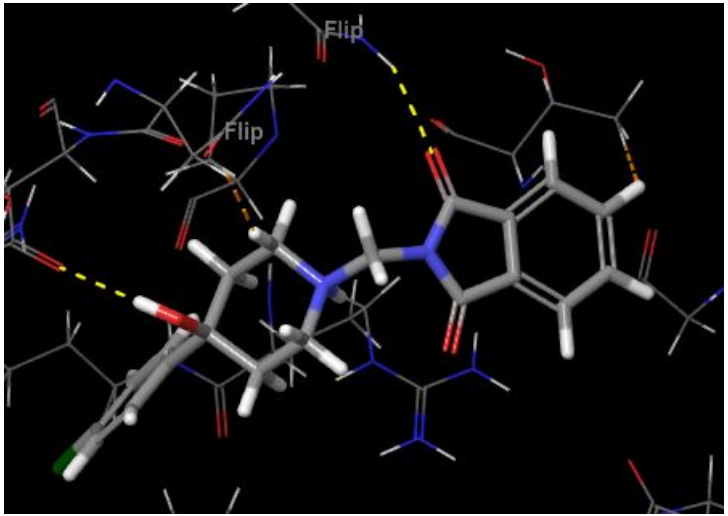
Intermolecular interactions: 3 hydrogen bonds and 2 bad contacts

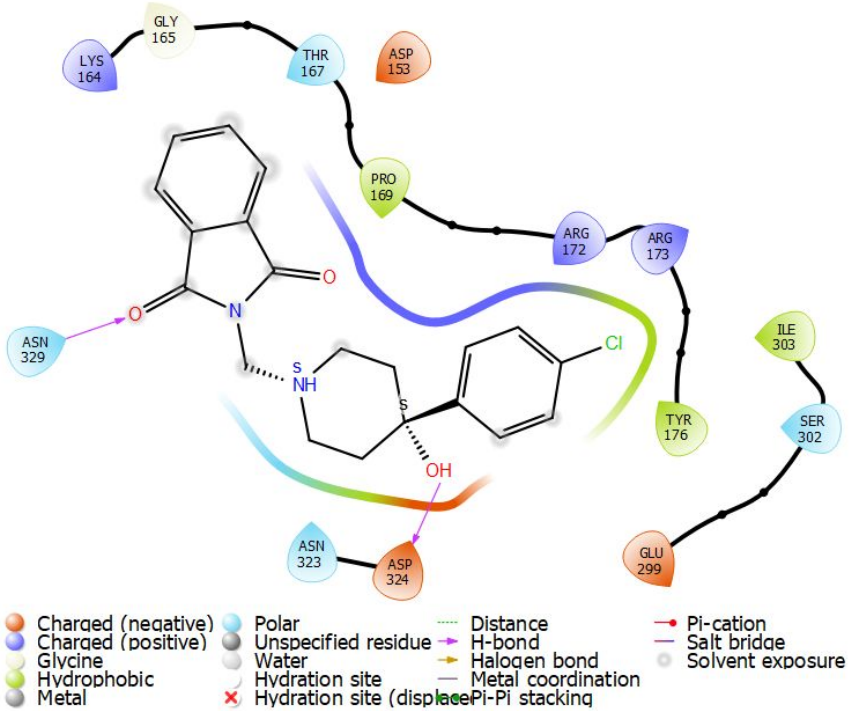




HTS 00735

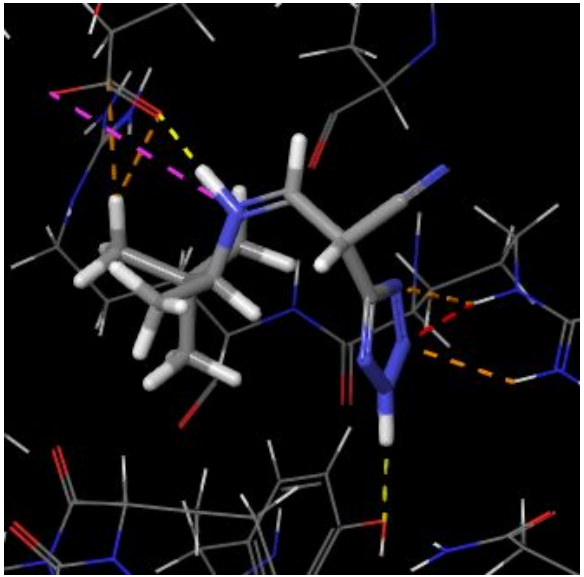
Intermolecular interactions: 2 hydrogen bonds and 2 bad contacts

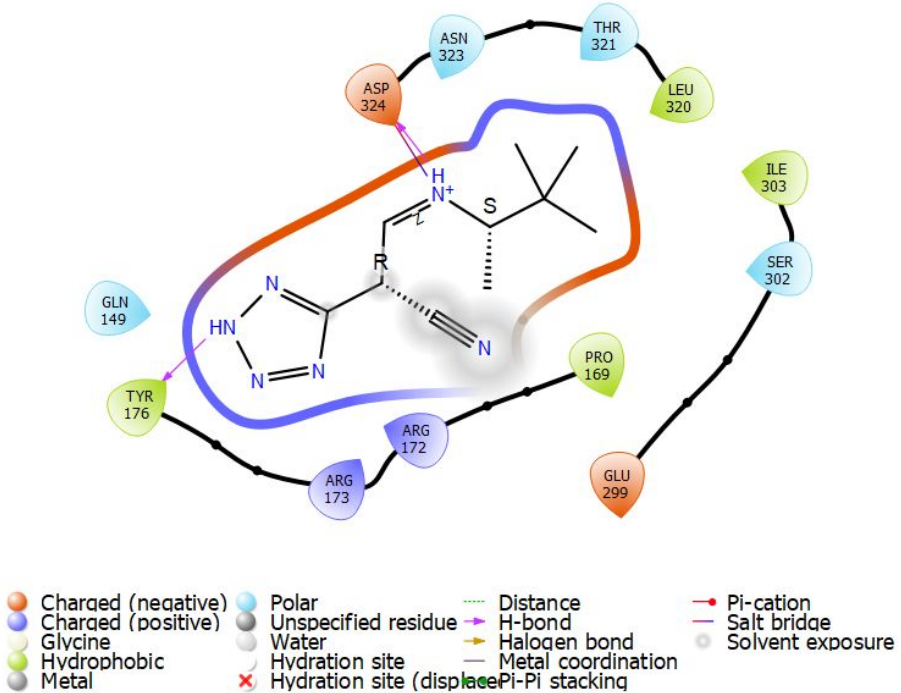




CD 11595

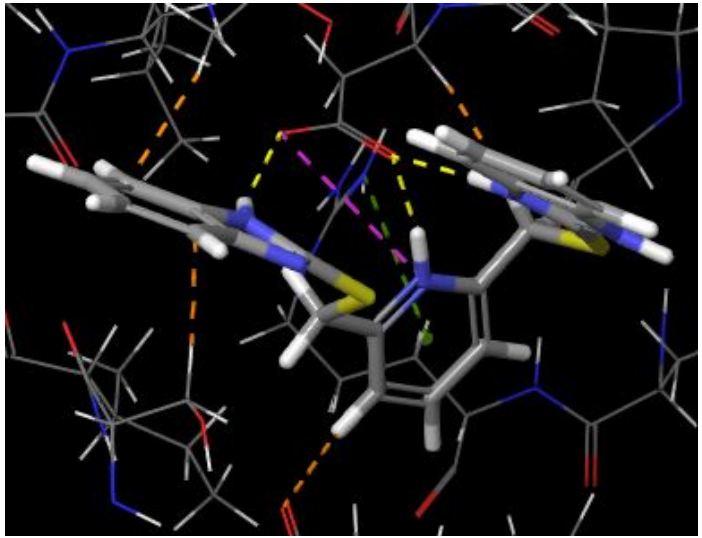
Intermolecular interactions: 2 hydrogen bonds, 1 salt bridge, 4 bad contacts and 1 ugly contact

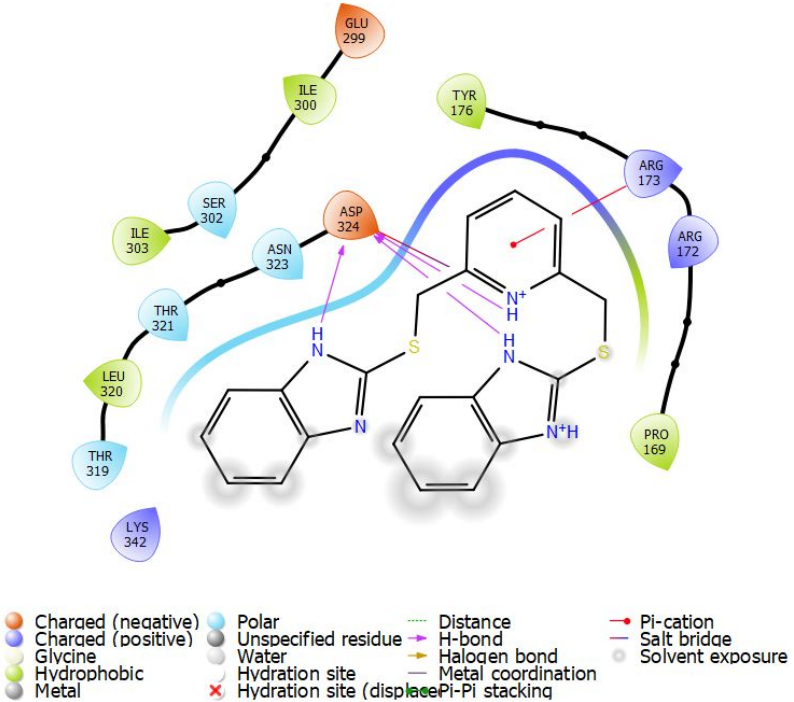




BTB 14890

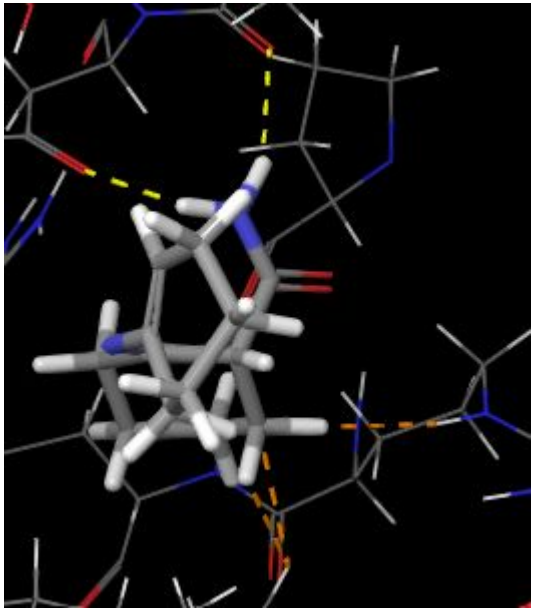
Intermolecular interactions: 3 hydrogen bonds, 1 salt bridge, 1 cation-π interaction and 4 bad contacts

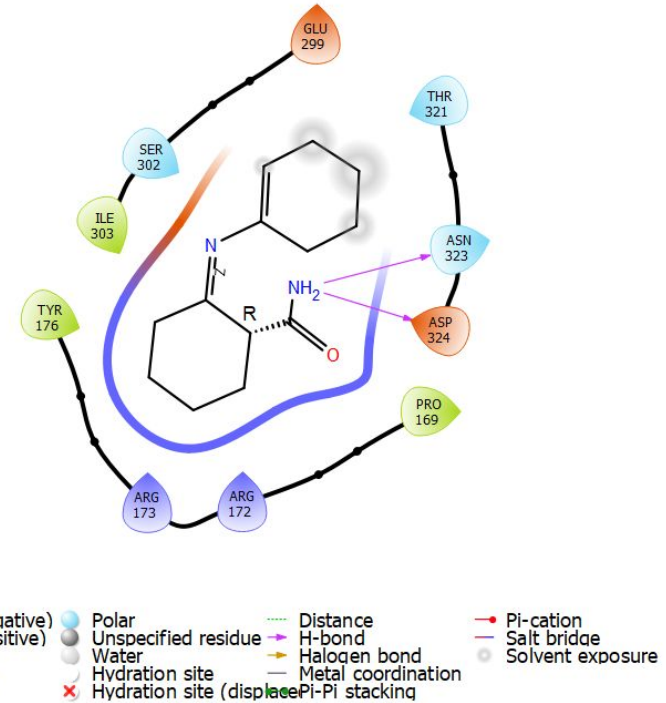




JFD 03877

Intermolecular interactions: 2 hydrogen bonds and 3 bad contacts





Closing Remarks

- The quality of an exemplar in intermolecular interactions with a protein depends on the specific goals of the drug design study and the properties of the protein.
- More intermolecular interactions with the protein suggest a stronger binding affinity.
- Solvent exposure can indicate a less stable or active compound, but in some cases, it can be desirable.
- Molecular docking techniques predict the binding mode of the molecule with the protein based on the 3D structures of both, and the output is a score that reflects the affinity of the molecule for the protein.
- The scores can be used to rank the exemplar compounds and choose the ones with the highest affinity for further evaluation and testing.
- Molecular dynamics simulations can study the dynamics of the molecule-protein complex and assess the stability of the binding.

Acknowledgements

I would like to thank Ms. Zoe Cournia, Alexios Chatzigoulas and Danai Kotzampasi for their teaching and support throughout the course.

Thank you for your attention.

References

1. 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.
2. Kotzampasi, Danai Maria, et al. "The orchestrated signaling by PI3Kα and PTEN at the membrane interface." Computational and Structural Biotechnology Journal (2022).
3. Chen, Chien-Yu, et al. "PTEN: tumor suppressor and metabolic regulator." Frontiers in endocrinology 9 (2018): 338.
4. Sansal, Isabelle, and William R. Sellers. "The biology and clinical relevance of the PTEN tumor suppressor pathway." Journal of clinical oncology 22.14 (2004): 2954-2963.
5. Salmena, Leonardo, Arkaitz Carracedo, and Pier Paolo Pandolfi. "Tenets of PTEN tumor suppression." Cell 133.3 (2008): 403-414.
6. Simpson, Laura, and Ramon Parsons. "PTEN: life as a tumor suppressor." Experimental cell research 264.1 (2001): 29-41.
7. Gkeka, P., & Cournia, Z. (2020, October). Molecular Dynamics simulations of lysozyme in water. Biomedical Research Foundation, Academy of Athens. Molecular Modeling of Biomolecules course, DSIT 2022-2023.
8. Cournia, Z. (2023). Principles of Computer-Aided Drug Design [PowerPoint presentation]. Molecular Modeling of Biomolecules course, DSIT 2022-2023.
9. Schrödinger, Maestro 10.2 User Manual, 2015, PDF.
10. Schrödinger, SiteMap 3.5 User Manual, 2015, PDF.
11. Schrödinger, Glide 6.7 User Manual, 2015, PDF.
12. Schrödinger, QikProp 4.4 User Manual, 2015, PDF.
13. 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.